A Novel Deep Learning Framework for Conditioned Feature Modulation and Synthetic Image Generation for Severity-Aware Disease Classification

MS (Research) Thesis

By

Saransh Vashistha



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

JULY 2026

A Novel Deep Learning Framework for Conditioned Feature Modulation and Synthetic Image Generation for Severity-Aware Disease Classification

A THESIS

submitted to the

INDIAN INSTITUTE OF TECHNOLOGY INDORE

in fulfillment of the requirements for

the award of the degree

of

MS (Research)

By

Saransh Vashistha



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

JULY 2026

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **"A Novel Deep Learning Framework for Conditioned Feature Modulation and Synthetic Image Generation for Severity-Aware Disease Classification"** in the fulfilment of the requirements for the award of the degree of MASTER OF SCIENCE (RESEARCH) and submitted in the DISCIPLINE OF COMPUTER SCIENCE AND ENGINEERING, Indian Institute of Technology Indore, is an authentic record of my own work carried out during the time period from August 2024 to May 2026 under the supervision of Prof. Aruna Tiwari, Professor, Indian Institute of Technology Indore, India.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

17 July, 2026

Signature of the Student with Date

(Saransh Vashistha)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

17/7/2026

Signature of Thesis Supervisor

with Date

(Prof. Aruna Tiwari)

ACKNOWLEDGEMENT

I owe a deep debt of gratitude to my thesis supervisor, **Prof. Aruna Tiwari**, whose intellectual rigour, patient guidance, and generous availability have been the constant backdrop to this work. She has a rare ability to see both the forest and the trees, pointing out the broader significance of an idea while simultaneously catching the error in an equation. I could not have asked for a better mentor.

I am equally grateful to **Dr. Milind Ratnaparkhe** and the scientists at the **ICAR-National Soybean Research Institute (ICAR-NSRI)**, Indore, who shared their botanical expertise so willingly. Their domain knowledge made the soybean severity annotations in this thesis far more meaningful than any algorithm alone could have. I also thank **AgriHub** for providing access to the AgriEdgeX computing infrastructure that enabled the large-scale generative experiments.

I am thankful to my research committee members for taking out valuable time to evaluate my progress throughout the course. Their helpful comments and suggestions improved this work at various stages. I am also grateful to **Dr. Ranveer Singh**, HOD of Computer Science and Engineering, for his continued help and support.

To my labmates – **Mr. Abhishek Tripathi**, **Ms. Sujata Gudge**, **Ms. Shilpy Kaur**, **Ms. Sharmila S.P.**, **Mr. Suhas Hegde**, and **Mr. Kishan Yadav** – thank you for the late-night debates, the readiness to listen whenever a result looked suspicious, and for making the lab a place worth coming back to every day. The intellectual honesty and warmth of that environment shaped me as a researcher more than I can adequately express.

My deepest gratitude goes to my family for their unwavering love and support throughout this journey. Their encouragement during the most challenging times kept me going. And to **Kaveri Gupta** – thank you for your patience, your belief in me, and for being there through every high and every setback.

Saransh Vashistha

To my family and friends

Abstract

Deep learning systems have achieved substantial progress in visual recognition and image synthesis, but two structural limitations continue to restrict their use in fine-grained, data-scarce settings. First, multitask recognition models often predict related outputs independently from a shared representation, even when one prediction should condition the interpretation of another. Second, generative models used for data augmentation often rely on class labels or generic text prompts, which provide limited information about the fine-grained visual structure of the target category. These limitations are especially important when labelled data are imbalanced and when the target labels have an ordered severity structure.

This thesis addresses these limitations through two complementary frameworks. The first, **Disease-Conditioned Severity Calibration (DCSC)**, is a multitask recognition framework that conditions a secondary prediction pathway on a co-predicted primary label. DCSC uses a VMamba backbone, FiLM-based feature modulation, disease-conditioned severity prototypes, CORAL ordinal regression, and an Earth Mover's Distance smoothing loss to make severity prediction context-aware rather than independent of the primary class prediction.

The second, **Knowledge-Grounded Depth-Controlled Diffusion (KGDC-Diffusion)**, is a two-stage generative framework for targeted synthetic data augmentation. It replaces simple class-label conditioning with retrieved morphological descriptors and uses pseudo-depth maps to impose spatial structure during generation. This allows the generative model to produce images that are not only visually realistic but also aligned with fine-grained category-specific visual attributes.

The proposed frameworks are evaluated on severity-aware plant disease diagnosis, which provides a practical testbed for both problems: severity labels depend on disease identity, and early-stage categories are underrepresented in field-collected datasets. On AI Challenger 2018, DCSC achieves 98.51% disease accuracy and 92.30% severity accuracy under a five-seed evaluation protocol, reducing soft severity MAE by 57% over the VMamba baseline. KGDC-Diffusion achieves an FID of 23.33, outperform-

ing the evaluated GAN and diffusion baselines, and improves downstream classifier accuracy by up to 3.76 percentage points when used for targeted augmentation of underrepresented classes. This thesis also curates **SoySeverityDataset**, a real-world, expert-validated soybean leaf image dataset with fine-grained severity annotations.

Overall, the thesis shows that making conditional structure explicit can improve both recognition and generation. Co-predicted labels can provide useful context for severity-aware prediction, while retrieved morphological knowledge can provide useful context for synthetic image generation in data-scarce, fine-grained visual domains.

Keywords: Multitask Learning, Feature-wise Linear Modulation, Ordinal Regression, Diffusion Models, Retrieval-Augmented Generation, ControlNet, Severity-Aware Classification, Synthetic Data Augmentation.

List of Publications

In Refereed Conferences

Saransh Vashistha and Aruna Tiwari. *Disease-Conditioned Severity Calibration for Plant Disease Multitask Learning*, Submitted to the International Conference on Neural Information Processing (ICONIP), 2026. (Rank B)

In Refereed Journals

Saransh Vashistha, Milind Ratnaparkhe, and Aruna Tiwari. *Knowledge-Grounded Depth-Controlled Diffusion for Synthetic Augmentation of Plant Disease Imagery*, submitted to Computers and Electronics in Agriculture, Elsevier

List of Abbreviations and Acronyms

DCSC Disease-Conditioned Severity Calibration

RAG Retrieval-Augmented Generation

KGDC-Diffusion Knowledge-Grounded Depth-Controlled Diffusion

FiLM Feature-wise Linear Modulation

CORAL Consistent Rank Logits

EMD Earth Mover's Distance

CNN Convolutional Neural Network

ViT Vision Transformer

Swin Shifted Window Transformer

SSM State Space Model

VSS Visual State Space

VMamba Visual State Space Model with Selective Scan

TSTC Triple-branch Swin Transformer Classification

LDI-NET Leaf Disease Identification Network

PD^{2}SE-Net Plant Disease Diagnosis and Severity Estimation Network

DA Disease Accuracy

SA Severity Accuracy

MAE Mean Absolute Error

GAN Generative Adversarial Network

DDPM Denoising Diffusion Probabilistic Model

LDM Latent Diffusion Model

VAE Variational Autoencoder

ADM-G Ablated Diffusion Model with Guidance

CLIP Contrastive Language–Image Pre-training

Min-SNR Minimum Signal-to-Noise Ratio

FID Fréchet Inception Distance

KID Kernel Inception Distance

IS Inception Score

YMV Yellow Mosaic Virus

AB Aerial Blight

BP Bacterial Pustule

ICAR-NSRI Indian Council of Agricultural Research – National Soybean Research Institute

Glossary of Plant Pathology Terms

Plant morphology The visible structure and form of a plant or plant organ, including leaf shape, venation, surface texture, colour pattern, and spatial arrangement of symptoms.

Symptom morphology The visual structure of disease symptoms, including their shape, size, colour, texture, distribution, and boundary characteristics.

Morphological descriptor A text description of visible disease characteristics, such as lesion shape, colour, margin, spread pattern, chlorosis, necrosis, or tissue damage.

Lesion A localised damaged or diseased region on plant tissue, often appearing as a spot, patch, necrotic area, or discoloured region on the leaf surface.

Lesion margin The boundary between diseased and healthy tissue, which may appear sharp, diffuse, irregular, chlorotic, or necrotic.

Chlorosis Yellowing of plant tissue caused by reduced or disrupted chlorophyll, commonly observed around infected or stressed leaf regions.

Necrosis Death of plant tissue, typically visible as brown, black, dry, or collapsed areas on the leaf.

Mottling An irregular pattern of light and dark discolouration on the leaf surface, often associated with chlorophyll disruption or viral infection.

Mosaic pattern A patchy arrangement of green and yellow regions on a leaf, commonly associated with viral disease symptoms.

Pustule A small raised blister-like structure on the leaf surface, often associated with bacterial or fungal infection.

Blight A symptom pattern involving rapid and extensive browning, wilting, or death of plant tissue.

Severity The degree or extent of disease expression on the plant, typically measured by the amount, spread, or intensity of visible symptoms.

Disease progression The temporal development of disease symptoms from early localised infection to advanced or widespread tissue damage.

Fine-grained disease classification Classification between visually similar disease categories or severity stages that differ by subtle symptom-level cues.

Contents

Abstract i

List of Publications iii

List of Abbreviations and Acronyms v

Glossary of Plant Pathology Terms vii

List of Figures xiii

List of Tables xvii

1 Introduction 1

1.1 Background 2

1.2 Motivation 4

1.3 Objectives 6

1.4 Thesis Contributions 7

1.5 Organisation of the Thesis 8

2 Related Work 11

2.1 Vision Backbone Architectures 11

2.1.1 Convolutional Neural Networks 12

2.1.2 Vision Transformers 13

2.1.3 The Swin Transformer 14

2.1.4 State Space Models and Visual State Space Models (VMamba) . 16

2.1.5 What VMamba Lacks: Cross-Task Conditioning 19

2.2 Multitask Learning, Feature Conditioning, and Ordinal Supervision . . 19
2.2.1 Multitask Learning in Computer Vision 20
2.2.2 Feature-wise Linear Modulation (FiLM) 21
2.2.3 Conditional Multitask Learning and Prior Cross-Task Work . . 22
2.2.4 Ordinal Regression and Consistent Rank Logits 23
2.3 Plant Disease Classification and Severity Estimation 24
2.3.1 Single-Task Plant Disease Classification 25
2.3.2 Multitask Disease-Severity Models 26
2.4 Generative Modelling for Data Augmentation 27
2.4.1 Generative Adversarial Networks 28
2.4.2 Latent Diffusion Models and Spatial Conditioning 29
2.4.3 The Conditioning Insufficiency Problem 31
2.4.4 Retrieval-Augmented Generation as a Conditioning Strategy . . 32

3 Visual State Space Model for Disease-Conditioned Severity Identification 34

3.1 Introduction 34
3.2 Proposed Methodology 35
3.2.1 Architecture Overview 36
3.2.2 FiLM-Based Disease-Conditioned Severity Modulation 38
3.2.3 CORAL Ordinal Severity Head 39
3.2.4 Disease-Conditioned Severity Prototypes 40
3.2.5 Earth Mover's Distance Loss for Distributional Smoothing . . . 41
3.2.6 Combined Loss and Training 42
3.3 Experimental Results 43
3.3.1 Datasets and Implementation 43
3.3.2 Baselines and Evaluation Metrics 45
3.3.3 Results on AI Challenger 2018 46
3.3.4 Results on Rice Leaf Severity 47
3.3.5 Statistical Significance 48

3.4 Ablation and Analysis . . . 48
3.4.1 Component Ablation . . . 48
3.4.2 Severity Confusion Matrices . . . 50
3.4.3 Feature Space Analysis . . . 51
3.4.4 Per-Disease Error and Uncertainty Quantification . . . 52
3.5 Summary . . . 52

4 Knowledge-Grounded Depth-Controlled Diffusion for Synthetic Augmentation of Plant Disease Imagery 54

4.1 Introduction . . . 54
4.2 Real Dataset Collection and Preprocessing . . . 56
4.2.1 SoySeverityDataset . . . 57
4.2.2 Botanical Text Corpus . . . 57
4.2.3 Pseudo-Depth Estimation . . . 59
4.3 Proposed Methodology . . . 60
4.3.1 Stage I: RAG-Guided Soft Semantic Conditioning . . . 62
4.3.2 Stage II: Depth-Controlled Spatial Conditioning . . . 64
4.4 Synthetic Generation and Targeted Augmentation Protocol . . . 66
4.4.1 Step A: Severity-Aware Generation . . . 67
4.4.2 Step B: Targeted Augmentation Construction . . . 67
4.5 Experimental Settings . . . 70
4.5.1 Datasets and Splits . . . 71
4.5.2 Implementation Details . . . 71
4.5.3 Generative Baselines . . . 71
4.5.4 Downstream Classifier Training . . . 73
4.5.5 Evaluation Metrics . . . 73
4.6 Results . . . 74
4.6.1 Generative Image Quality . . . 75
4.6.2 Downstream Classification on SoySeverityDataset . . . 79
4.7 Ablation and Sensitivity Analysis . . . 82

4.7.1 Prompt Conditioning Ablation . . . 83
4.7.2 Component Ablation . . . 83
4.7.3 Guidance-Scale Sensitivity . . . 84
4.7.4 Depth-Conditioning Scale Sensitivity . . . 85
4.8 Computational Cost and Practical Feasibility . . . 86
4.9 Cross-Dataset Portability . . . 88
4.9.1 Controlled-Environment Portability: PlantVillage Tomato . . . 88
4.9.2 Real-Field Generative Portability: CottonWeedID10 and DeepWeeds . . . 89
4.9.3 Cross-Dataset Downstream Validation: CottonWeedID15 . . . 90
4.10 Summary . . . 92

5 Conclusion and Future Work **94**
5.1 Summary of Contributions . . . 94
5.2 Limitations . . . 96
5.3 Future Work . . . 96

Bibliography **97**

List of Figures

2.1 CNN (a) versus Vision Transformer (b): local versus global feature aggregation. 13

2.2 Swin Transformer shifted window attention mechanism. 15

2.3 VMamba's SS2D module: four-directional selective scan for omnidirectional spatial context. 18

2.4 Feature-wise Linear Modulation (FiLM) block. 22

2.5 DDPM forward and reverse diffusion processes. 29

2.6 ControlNet adapter architecture for spatial conditioning of Latent Diffusion Models. 30

3.1 DCSC architecture. The shared VMamba-Small backbone feeds a disease head and a severity projection. The predicted disease identity drives a FiLM generator that modulates the severity feature space. CORAL cumulative outputs are converted to a three-class distribution and fused with the disease-conditioned prototype distribution. Training uses disease cross-entropy, CORAL, prototype, and Earth Mover's Distance objectives. 37

3.2 Representative samples from AI Challenger 2018 illustrating the disease-conditional nature of severity criteria. The visual boundary between general and serious severity varies substantially across disease types, which motivates the disease-conditioned modulation mechanism in DCSC. 44

3.3 Representative fixed-seed severity confusion matrices for the reproduced TSTC baseline (left) and DCSC (right) on AI Challenger 2018. DCSC reduces serious-to-general errors from 335 to 158 (52.84%); general-to-serious errors change from 187 to 185. Table 3.1 reports five-seed means. 50

3.4 UMAP projections of severity features for all AI Challenger 2018 test samples, before and after FiLM modulation. The post-FiLM general-to-serious centroid distance increases by 34% (from 0.570 to 0.766) and the mean within-disease general-to-serious separation increases by 40% (from 0.722 to 1.014), confirming that FiLM creates disease-specific severity subspaces as designed. 51

4.1 Class-wise distribution of the SoySeverityDataset across disease categories and severity levels, highlighting the natural long-tail class imbalance present in field-collected samples. 58

4.2 High-level block diagram depicting the overall KGDC-Diffusion framework. Stage I (RAG-guided semantic conditioning) and Stage II (depth-controlled spatial conditioning) are each trained once and then kept frozen. Steps A–C describe the synthesis, augmentation construction, and downstream evaluation protocol. 61

4.3 Illustration of Stage I of the proposed pipeline. Section (A) details the prompt orchestration via Retrieval-Augmented Generation (RAG); and Section (B) demonstrates the semantic fine-tuning of the Latent Diffusion Model. 62

4.4 Stage II depth conditioning: a pseudo-depth map estimated from the RGB image conditions a depth ControlNet attached to the frozen, semantically fine-tuned Stable Diffusion backbone, providing coarse structural guidance over leaf geometry during generation. 65

4.5 Representative pseudo-depth maps (from DepthAnything) alongside original and generated images under three augmented depth conditions (15° rotation, 45° rotation, and horizontal flip). Beyond geometric re-orientation, each augmented depth map drives the denoising process along a distinct trajectory, producing outputs that differ in illumination direction, colour temperature, lesion shading, and background context, photometric diversity that is critical for downstream classifier generalisation. 66

4.6 Qualitative comparison of synthetic images across representative disease–severity categories (AB: Aerial Blight, BP: Bacterial Pustule, YMV: Yellow Mosaic Virus; severity levels: Early, Mild, Late) alongside real samples. Each column corresponds to one model; each row corresponds to one disease–severity class. 76

4.7 Class-stratified UMAP projections of Inception-V3 features, separating early-stage (top row) from mild/late (bottom row) disease categories across evaluated models. 78

List of Tables

3.1 Five-seed test results on AI Challenger 2018 (70/15/15 fixed split). Bold indicates the best value in each column. DA = disease accuracy, SA = hard severity accuracy, Comb = unweighted mean of DA and SA, and Soft MAE is computed from the expected severity. 46

3.2 Five-seed test results on the Rice Leaf Severity dataset. Bold indicates the best value in each column. Disease accuracy at ceiling for several methods isolates the severity contribution. 47

3.3 Paired t-test results on AI Challenger 2018 across five seeds. Asterisk indicates $p < 0.05$. 48

3.4 Five-seed component ablation study on AI Challenger 2018 using the fixed 70/15/15 train–validation–test split. DA and SA denote disease accuracy and severity accuracy, respectively. The best result in each column is shown in bold. 49

4.1 Representative entries from the botanical knowledge corpus (two per class shown). The full corpus of 80 entries is provided in the supplementary material. 59

4.2 KGDC-Diffusion inference procedure (Step A). 68

4.3 Per-class real-data split and realism-filtered synthetic additions for the SoySeverityDataset. The balancing target is $T = n_{\max} = 489$ (Healthy class); 800 candidate images are generated for every class with $n_c < T$ and then filtered using the realism-score criterion of Eq. 4.4. 70

4.4 Exhaustive list of hyperparameters used for Stage I (Base LDM) and Stage II (Depth ControlNet) training. 72

4.5 Quantitative comparison of images generated by different generative architectures. FID, KID, and IS are computed over the complete generated sets. Precision and Recall are computed using the dataset-trained VGG16 feature extractor. Arrows indicate the desired direction of improvement. The best result in each column is highlighted in **bold**. . . . 77

4.6 Per-class generative quality measured by KID on the SoySeverity-Dataset train split, using subset size $\min(50, N_c)$. 79

4.7 Downstream classification performance (mean ± standard deviation over five retained seeds) for each architecture trained without and with realism-filtered targeted synthetic augmentation. Precision, Recall, and F1 are macro-averaged over the ten disease–severity classes. Bold in the Δ rows marks the largest gain in each metric across architectures. . . . 81

4.8 Paired t-test results for classification accuracy (original vs. augmented), $n = 5$ retained seeds per model. 81

4.9 Class-wise classification performance of ResNet-50 before and after realism-filtered targeted synthetic data augmentation, averaged over five retained seeds. 82

4.10 Prompt-conditioning ablation on the SoySeverityDataset at a common reference guidance scale. Class-name is the fine-tuned SD baseline; retrieval is evaluated for $k = 1$ to 6. 83

4.11 Component ablation of the KGDC-Diffusion framework. Each row adds one component to the previous configuration. Metrics are computed against the training-set reference (Inception-V3 for FID/KID; dataset-trained VGG16 for Precision/Recall). 84

4.12 Guidance-scale sensitivity for the retrieval-grounded model (without depth ControlNet) on the SoySeverityDataset, evaluated against the training-set reference. Precision and Recall are computed in the dataset-trained VGG16 feature space. 85

4.13 Depth ControlNet conditioning-scale sensitivity (guidance scale fixed at 3.0). Precision and Recall are computed in the dataset-trained VGG16 feature space; ± values indicate variation across repeated synthetic sampling/evaluation runs. 86
4.14 Inference compute cost per model at 512×512, batch size 8. FLOPs for diffusion models are total over all sampling steps. 87
4.15 Training compute cost per optimiser step at 512×512, batch size 8. ADM rows reflect a single generative-model step. The SD1.5 ControlNet row is ControlNet-only (U-Net frozen). 87
4.16 External validation on the PlantVillage tomato subset. The depth-controlled model substantially improves over the fine-tuned class-name baseline. 89
4.17 Generative performance on the CottonWeedID10 weed-species dataset. The best value in each column is highlighted in **bold**. 89
4.18 Generative performance on the DeepWeeds weed-species dataset. The best value in each column is highlighted in **bold**. 90
4.19 Downstream classification performance on CottonWeedID15 (mean ± standard deviation over five seeds) for each architecture trained without and with filtered targeted synthetic augmentation. 91
4.20 Paired t-test and one-sided Wilcoxon signed-rank test results for CottonWeedID15 classification accuracy (original vs. filtered-augmented), $n = 5$ seeds per model. 92

Chapter 1

Introduction

The past decade has witnessed a fundamental transformation in computer vision, driven by deep neural networks that extract increasingly rich hierarchical representations from raw pixel data. This progress has opened the door to practical automated systems for visual recognition and image synthesis across a wide range of specialised domains, from medical imaging to precision agriculture. Yet despite rapid advances in backbone architectures and generative models, two structural problems persist across both recognition and synthesis: *how to condition one prediction on another in multi-task learning*, and *how to generate fine-grained, domain-faithful images when training data is scarce*. This thesis addresses both problems in the context of plant disease recognition, a domain with a natural hierarchical label structure, established benchmarks, and a well-documented long-tail data distribution.

The first problem matters because co-occurring labels such as disease type and severity are not independent. The visual criteria for severity differ fundamentally across disease identities, and ignoring this dependency produces systematic classification errors that no amount of backbone scaling is likely to resolve. The second problem matters because the classes most critical for timely intervention such as early-stage infections are precisely those most underrepresented in field-collected datasets, and existing generative conditioning strategies lack the morphological specificity needed to synthesise them faithfully. Taken together, the thesis frames both contributions as a study of *conditioning*: conditioning recognition decisions on co-predicted labels, and

conditioning generative synthesis on expert domain knowledge.

1.1 Background

Visual recognition has evolved through successive generations of backbone architectures, each addressing limitations of its predecessor. Early convolutional neural networks (CNNs) demonstrated that hierarchical feature learning from raw pixels could surpass hand-crafted feature pipelines on large-scale benchmarks (Ferentinos, 2018). Deeper and wider convolutional architectures improved representational capacity but remained limited by their local receptive fields, constraining the modelling of long-range spatial dependencies (Chen et al., 2020).

Vision Transformers (ViTs) addressed this limitation by replacing local convolution with global self-attention over image patches, allowing non-local feature interactions throughout the image (Dosovitskiy et al., 2021). The Shifted Window Transformer (Swin) (Liu et al., 2021) introduced a hierarchical patch-merging strategy that recovered the multi-scale inductive bias of CNNs while retaining long-range attention, establishing state-of-the-art performance on image classification, object detection, and semantic segmentation. However, the core cost of self-attention remains quadratic in the number of tokens, which limits scalability at high spatial resolutions.

State Space Models (SSMs) offer a computationally efficient alternative rooted in classical control theory. The S4 model (Gu et al., 2022) parameterised the SSM state matrix using the HiPPO polynomial projection framework, enabling parallel training via convolutional kernels and achieving superior performance on long-range dependency benchmarks over all prior Transformer, RNN, and CNN alternatives. Mamba (Gu and Dao, 2024) extended this by introducing a selective scan mechanism (S6), in which the SSM parameters are computed as functions of the input, enabling content-based selective propagation at linear computational cost. VMamba (Liu et al., 2024) adapted this selective scan to two-dimensional image data through a 2D Selective Scan module (SS2D) that traverses feature maps along four directions, yielding a hierarchical vision backbone structurally analogous to Swin but with linear computational

complexity and superior performance across standard vision benchmarks.

Multitask learning in visual recognition addresses settings where multiple related outputs, such as disease type and severity grade, must be predicted simultaneously from a shared input. The dominant paradigm is hard-parameter-sharing, where a single backbone feeds independent task-specific heads from a common representation (Crawshaw, 2020). This architecture is efficient and acts as a regulariser, but it treats all tasks as independent and discards the conditional dependency structure that naturally exists among co-occurring labels. When one label's visual criteria depend fundamentally on the value of another, a shared backbone with independent heads cannot encode this dependency, because there is no pathway for one task's prediction to influence the feature computation of another. Despite the representational advances of VMamba and its predecessors, this structural limitation persists across all existing visual standard shared backbone architectures: none provides a mechanism for conditioning one task's feature space on the output of a co-predicted task.

In parallel, generative modelling for image synthesis has undergone an equally dramatic transformation. Generative Adversarial Networks (Goodfellow et al., 2014) established adversarial training as a framework for high-fidelity image synthesis, but they suffer from training instabilities and mode collapse in low-data regimes. Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) provided a more stable generative framework based on learned score matching, achieving superior sample diversity. Latent Diffusion Models (LDM) (Rombach et al., 2022) moved the diffusion process into the compressed latent space of a variational autoencoder, reducing computational cost substantially while preserving generation quality. ControlNet (Zhang et al., 2023) further introduced spatially precise conditioning via trainable adapter layers, enabling segmentation-mask, depth-map, and edge-map guidance. However, most existing generative approaches rely on discrete class labels or open-domain text prompts as conditioning signals. These signals may be effective in large-scale natural-image settings, where models are exposed to millions of training examples, but they are often insufficient in specialised scientific domains. A class label only specifies the target category and does not describe its morphology, such as lesion shape, spatial

spread, colour variation, or structural appearance. Similarly, generic text prompts may lack the domain-specific grounding needed to capture fine-grained visual details.

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) was developed in natural language processing to ground model outputs in dynamically retrieved, domain-specific factual context, offering a route to richer conditioning signals without retraining the base generative model. Feature-wise Linear Modulation (FiLM) (Perez et al., 2018) and ordinal regression frameworks such as (COnsistent RAnk Logits) (Cao et al., 2020) provide the tools for conditioning and supervising cross-task predictions in the recognition setting. Together, these methods offer targeted solutions to the two gaps identified above.

Hence, both lines of development share a common unresolved limitation. Visual backbone architectures, regardless of their representational strength, treat co-predicted tasks as independent and provide no mechanism for one task's decision to be informed by another's output. Generative models, regardless of their synthesis quality, condition generation on class indices or generic text strings that carry no information about what the target class actually looks like morphologically. In both cases, the conditional structure that governs the domain is absent from the architecture. We propose to investigate whether making this conditional structure explicit, by conditioning severity recognition on predicted disease identity and conditioning generative synthesis on retrieved expert morphological descriptions, can resolve these limitations in the context of plant disease analysis.

1.2 Motivation

Over the past decade, visual recognition backbones have progressed from convolutional networks to attention-based Transformers to, most recently, selective state-space models such as VMamba, each generation trading one limitation for another: convolutional locality, the quadratic cost of self-attention, and now the absence of any mechanism for one prediction to condition another within a shared backbone. Generative modelling has followed a parallel trajectory, from adversarial training through

diffusion-based synthesis to spatially precise conditioning via ControlNet, yet remains tied to conditioning signals, class indices or text prompts, that carry no morphological information about the target class. Hence, we propose to investigate whether making the conditional structure of the domain explicit, rather than treating it as incidental, can resolve both limitations.

A VMamba backbone with independent task heads inherits the linear-time efficiency of selective scanning, but it still treats disease identity and severity as independent outputs of a shared representation, even though the visual evidence for severity is not defined in isolation: the features indicating severe damage for one disease are not the features indicating severe damage for another. This gap is structural rather than representational, since no increase in the shared backbone's representational capacity alone can resolve it. The actual limitation lies in the task-independence assumption of the head structure, not in how much the backbone can represent. We therefore propose to investigate severity-aware plant disease diagnosis by adding a structural block that conditions the severity pathway on predicted disease identity through FiLM-based feature modulation, and by supervising it with an ordinal objective reflecting the natural ordering of severity grades.

Diffusion models conditioned on class indices or text prompts are only told which disease a leaf has, not what that disease actually looks like at each stage. They are not given the visual details, such as the shape of the affected area, its colour, or how it spreads across the leaf, that separate one disease's early appearance from another's. This is a gap in the conditioning signal provided to the model, not a limitation of the diffusion process itself. The problem is worse for early-stage infections precisely because they are the least represented cases in existing datasets: there are simply fewer early-stage images for the model to learn from in the first place.

We therefore propose to investigate whether retrieving expert descriptions of disease appearance at generation time, used to condition a ControlNet-guided diffusion process through a retrieval-augmented pipeline, can produce synthetic images faithful enough to improve classifier performance on these under-represented classes. Both investigations are instances of the same underlying question: whether resolving condi-

tioning gaps, conditioning one task's prediction on another's output, and conditioning generation on domain knowledge rather than a label index, yields measurable gains where structural independence currently limits both recognition and synthesis. Plant disease diagnosis offers a domain where this can be tested concretely and where the stakes are direct: misclassifying severity delays or misdirects treatment, and the inability to synthesise faithful early-stage images leaves exactly the cases that matter most for early intervention underserved by data augmentation.

1.3 Objectives

The two investigations proposed in Section 1.2, resolving the absence of cross-task conditioning in multitask recognition, and resolving the absence of morphologically grounded conditioning in generative synthesis, translate into three concrete objectives. The first two address the recognition gap identified in Section 1.1: that a VMamba backbone with independent task heads, regardless of its linear-time efficiency, treats disease identity and severity as unconditioned outputs of a shared representation. The third addresses the generative gap: that existing diffusion-based augmentation conditions synthesis on class indices or text prompts that carry no morphological information about the target class, precisely where such information matters most, namely in underrepresented early-stage classes. In this thesis, we aim to achieve the following objectives:

1. Enhance the VMamba vision backbone with a disease-conditioned cross-task feature modulation mechanism to enable severity-aware multitask plant disease learning. Specifically, develop a framework in which the severity prediction pathway adapts to the predicted disease identity through FiLM-based affine modulation, disease-conditioned severity prototypes, CORAL ordinal regression, and an Earth Mover's Distance smoothing loss, in order to address the systematic confusion between general and serious severity that arises from disease-agnostic decision boundaries.

2. Establish, through a rigorous multi-seed evaluation protocol (fixed 70/15/15 split, five independent seeds, mean ± standard deviation, and paired t-tests for statistical significance), that the gains attributable to disease-conditioned severity modulation are consistent and statistically significant, rather than artefacts of a single train/test split.

3. Develop a two-stage knowledge-grounded, depth-controlled generative pipeline that replaces class-label conditioning with expert-validated morphological descriptors and spatially precise pseudo-depth ControlNet supervision, producing biologically faithful early-stage disease and weed images that improve downstream classifier performance on underrepresented classes.

1.4 Thesis Contributions

The specific contributions of this thesis are as follows.

- **Disease-Conditioned Severity Calibration (DCSC).** We identify disease-conditional severity ambiguity as the primary structural bottleneck in severity-aware plant disease multitask learning: the visual boundary between general and serious severity differs across disease types, yet all prior methods impose a shared decision boundary. Conceptually, DCSC is the first framework to encode the insight that severity is not an absolute visual quantity but a disease-relative one, a principle that mirrors how trained plant pathologists actually perform severity assessment. By making this inductive bias explicit at the feature level rather than leaving the network to discover it implicitly, DCSC converts what was an inherent architectural limitation into a learnable, interpretable mechanism. DCSC addresses this by conditioning the severity prediction pathway on the predicted disease identity through complementary mechanisms: FiLM-based affine modulation of the severity feature space; a learnable disease-conditioned severity prototype matrix with ordinal geometric regularisation; a CORAL ordinal regression head enforcing rank-monotone predictions; and an Earth Mover's

Distance loss for ordinal distributional smoothing. Evaluated on AI Challenger 2018 under a rigorous five-seed 70/15/15 fixed-split protocol, DCSC achieves 98.51% disease accuracy and 92.30% severity accuracy with a soft severity MAE of 0.0909, representing a 57% soft-MAE reduction over the VMamba baseline and statistically significant improvements over both TSTC and LDI-NET.

- **Knowledge-Grounded Depth-Controlled Diffusion (KGDC-Diffusion).** We propose a two-stage generative pipeline for producing biologically faithful early-stage plant disease images. The major goal of this proposed work is to produce faithful and accurate synthetic images that lie in the same manifold and distributional space as the original data samples. Stage I uses Retrieval-Augmented Generation to replace vague class-label conditioning with expert-validated morphological descriptors retrieved at inference time from a domain-specific botanical knowledge base. Stage II conditions a ControlNet architecture on pseudo-depth maps from a zero-shot monocular depth estimator to enforce spatial and geometric realism. KGDC-Diffusion achieves an FID of 23.33 and a Generative Recall of 0.951, outperforming all GAN and diffusion baselines, and downstream classifiers augmented with KGDC-Diffusion synthetic data gain up to 3.76 percentage points in accuracy over classifiers trained without augmentation.

- **SoySeverityDataset.** We curate and release a real-world soybean leaf disease dataset validated by a domain expert at ICAR-NSRI, comprising three disease classes – Aerial Blight, Yellow Mosaic Virus, and Bacterial Pustule across three severity levels (Early, Mild, and Late). This dataset fills a gap in publicly available soybean disease benchmarks with fine-grained severity and spatial annotation.

1.5 Organisation of the Thesis

The remainder of this thesis is organised as follows.

Chapter 2 (Related Work)

This chapter reviews the literature across six areas that underpin the two contributions of the thesis. It covers the evolution of sequence modelling from structured state space models (S4) through Mamba's selective scan mechanism to VMamba's 2D adaptation, and identifies the missing cross-task conditioning capability. It then reviews multitask learning and FiLM-based feature conditioning, ordinal regression and severity grading, generative modelling from GANs through latent diffusion models, the limitations of existing conditioning strategies in specialised domains, and prior multitask plant disease classification work.

Chapter 3 (Disease-Conditioned Severity Calibration)

This chapter presents the DCSC framework in full detail. It covers the problem formulation, the FiLM-based conditioning module, the disease-conditioned severity prototype head with ordinal geometric regularisation, the CORAL ordinal severity head, the Earth Mover's Distance smoothing loss, and the combined training objective. Experiments on AI Challenger 2018 and the Rice Leaf Severity dataset include a comprehensive ablation study, statistical significance tests, per-disease severity error analysis, and Monte Carlo Dropout uncertainty analysis on the five highest-error disease classes.

Chapter 4 (Knowledge-Grounded Depth-Controlled Diffusion)

This chapter presents the KGDC-Diffusion generative framework. It describes the SoySeverityDataset collection, expert severity annotation, and botanical corpus construction, followed by the two-stage pipeline architecture comprising RAG-guided semantic conditioning and depth-controlled spatial conditioning via a pseudo-depth ControlNet. Experiments evaluate generation quality (FID, KID, Generative Precision and Recall) against GAN and diffusion baselines across four datasets, followed by downstream classification experiments demonstrating the augmentation benefit of targeted severity-aware synthetic data across three classifier architectures.

Chapter 5 (Conclusion and Future Work)

This chapter summarises the contributions of the thesis, discusses limitations of both frameworks, and outlines directions for future research including cross-attention-

based conditioning, infection-type-aware hierarchical priors, automatic loss weighting, and cross-domain validation.

Chapter 2

Related Work

This chapter builds the conceptual case for both contributions of the thesis by reviewing the prior work each contribution responds to. Section 2.1 traces the evolution of visual backbone architectures from convolutional networks through Vision Transformers and the Swin Transformer, arrives at the resolution-complexity bottleneck that limits attention-based backbones, and shows how VMamba's selective state space mechanism resolves that bottleneck while introducing a new gap: the absence of cross-task conditioning. Section 2.2 reviews multitask learning, feature conditioning, and ordinal supervision, explaining why FiLM combined with CORAL is the appropriate combination to fill the gap. Section 2.3 surveys automated plant disease classification and the multitask disease-severity literature to situate the thesis benchmarks and baselines. Section 2.4 reviews generative modelling approaches for data augmentation, traces the conditioning insufficiency problem that limits existing diffusion-based approaches in scientific domains, and justifies the retrieval-augmented strategy used in KGDC-Diffusion. Agricultural application contexts appear only where they motivate specific design choices.

2.1 Vision Backbone Architectures

Understanding why this thesis uses VMamba as its recognition backbone, and why that choice matters for the experimental comparison against prior multitask meth-

ods, requires tracing the lineage of visual backbone architectures. Each generation of backbone addressed a specific limitation of its predecessor, and understanding those limitations is the most direct way to understand what VMamba provides and what it still does not.

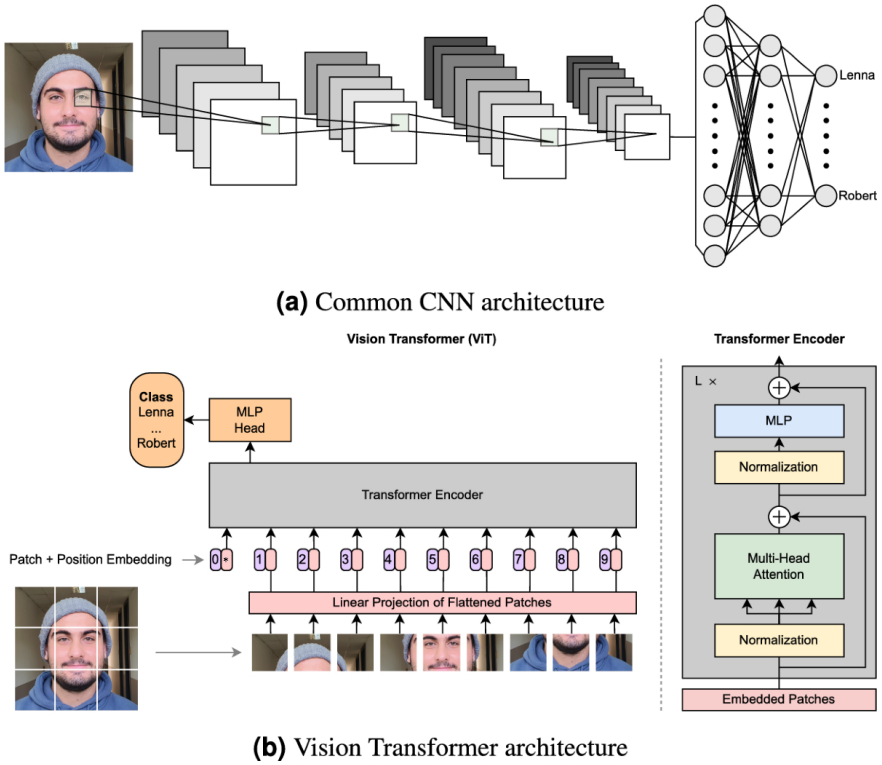
2.1.1 Convolutional Neural Networks

For most of the last decade, the convolutional neural network (CNN) was the dominant building block of visual recognition. A CNN applies a bank of learned filters across the spatial dimensions of an image in a sliding-window fashion, producing a feature map that encodes local image statistics at each position. The key insight behind the design, that natural images are spatially local because the pixels jointly defining an edge, a texture, or a lesion boundary tend to be neighbours, made CNNs enormously effective on image classification tasks. AlexNet (Krizhevsky et al., 2012) demonstrated this at scale on ImageNet, and subsequent architectures including VGGNet (Simonyan and Zisserman, 2015), ResNet (He et al., 2016), and DenseNet (Huang et al., 2017) progressively deepened and refined the convolutional design, achieving human-level performance on standard benchmarks.

In the plant disease domain specifically, CNNs pre-trained on ImageNet and fine-tuned on leaf disease datasets demonstrated strong diagnostic performance, with several works showing that CNN classifiers could match or approach trained pathologists on benchmark tasks (Ferentinos, 2018; Chen et al., 2020; Yang et al., 2020). This is not surprising given the task structure. A rust pustule, a bacterial lesion halo, or a powdery mildew spot are all local patterns that convolution is well-suited to detect.

The limitation of CNNs is that their receptive field, the region of the image that influences any single output activation, grows only linearly with depth. A 3x3 convolutional layer has a receptive field of 3x3, and stacking L such layers gives a theoretical receptive field of $(2L+1) \times (2L+1)$. In practice, due to subsampling and the decay of gradient signal through many layers, effective receptive fields are much smaller than the theoretical maximum (Luo et al., 2016). For a task like severity grading, where the overall distribution of lesions across the entire leaf surface matters as much as any

individual lesion, this local bias becomes a bottleneck. A CNN processing the tip of a leaf has limited access to information about what is happening near the stem, yet both regions together determine the overall severity grade. This is the gap that Vision Transformers were designed to fill.



(a) Common CNN architecture

(b) Vision Transformer architecture

Figure 2.1: CNN (a) versus Vision Transformer (b): local versus global feature aggregation.

2.1.2 Vision Transformers

The Transformer architecture was introduced for sequence transduction by Vaswani et al. (2017). Dosovitskiy et al. (2021) subsequently adapted the Transformer encoder to computer vision in the Vision Transformer (ViT). ViT divides an input image into a grid of fixed-size non-overlapping patches, linearly projects each patch into an embedding vector, and processes the resulting sequence of patch embeddings with standard Transformer self-attention layers. Self-attention allows every patch in the image to attend directly to every other patch regardless of spatial distance.

An image at resolution 224×224 divided into 16×16 patches produces $N = 196$

patch tokens. In each self-attention layer, the model computes a query, key, and value vector for each of the 196 tokens, and then computes 196 attention weights for each token by comparing its query against every other token's key. The attention weight between patch i and patch j tells the model how much patch i's representation should be updated based on patch j's content. Because this comparison is computed for every pair (i, j), the total number of attention computations is $N^2 = 38{,}416$ per attention head per layer. More generally, the computational and memory cost of self-attention scales as $O(N^2)$ in both time and space. This is acceptable at 224×224 but becomes prohibitive as resolution increases. At 448×448 with the same patch size, $N = 784$ and $N^2 = 614{,}656$, a sixteen-fold increase in attention cost for a four-fold increase in image resolution.

ViT-based models demonstrated that global self-attention does provide substantially better feature quality than CNNs on large-scale benchmarks, confirming that long-range spatial dependencies matter for visual recognition. However, ViT in its original form requires very large training datasets such as JFT-300M to outperform strong CNNs, partly because it discards the spatial inductive bias of convolution entirely and must learn all spatial relationships from data.

2.1.3 The Swin Transformer

The Swin Transformer (Liu et al., 2021) addressed both the quadratic scaling and the single-scale limitations of ViT through two design choices. First, Swin restricts self-attention to local non-overlapping windows rather than computing it globally. Each window contains $M \times M$ patches (typically $M = 7$), and attention is computed independently within each window. This reduces the attention complexity from $O(N^2)$ to $O(N \cdot M^2)$, which is linear in N for fixed M and thus far more tractable at higher resolutions.

Second, to allow information to flow across windows, since windowed attention alone would make patches from different windows permanently isolated, Swin alternates between a regular window partition and a shifted window partition in successive layers. The shift displaces the window boundaries by $(\lfloor M/2 \rfloor, \lfloor M/2 \rfloor)$ pixels, creat-

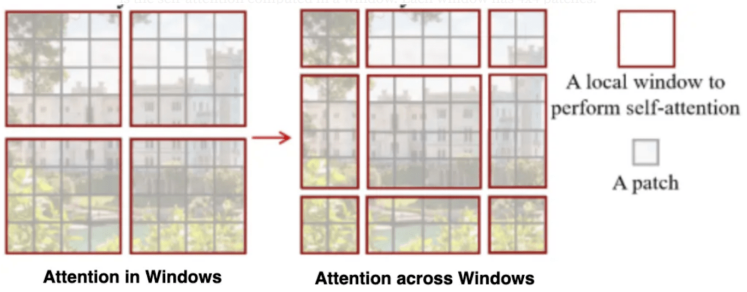


Figure 2.2: Swin Transformer shifted window attention mechanism.

ing new windows that overlap with the boundaries of the previous partition. This allows cross-window communication at every pair of layers, providing a form of global connectivity despite local computation. Figure 2.2 illustrates this mechanism.

Third, Swin organises its blocks into a four-stage hierarchical architecture with patch-merging layers between stages. Each patch-merging layer doubles the stride and halves the spatial resolution, producing a pyramid of feature maps at scales $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$. This multi-scale hierarchy is what allows Swin to excel at dense prediction tasks such as detection and segmentation in addition to image classification. On ImageNet classification, COCO detection, and ADE20K segmentation, Swin Transformer substantially outperformed all prior CNN and ViT models at the time of publication.

In the plant disease multitask literature, Swin Transformer was adopted as the backbone for TSTC (Yang et al., 2023), currently the strongest prior work on joint disease-severity prediction, achieving 99.00% disease accuracy and 88.73% severity accuracy on AI Challenger 2018. Despite this success, the Swin Transformer still carries the $O(N \cdot M^2)$ scaling with respect to the number of tokens. While M^2 is a constant, the total computational cost still grows linearly with N, and more critically, the window size M is fixed at architecture design time. Expanding M to capture larger spatial contexts immediately reintroduces quadratic cost within the window. This means that for high-resolution inputs or tasks requiring long-range spatial context, even Swin's windowed attention faces a hard tradeoff between context and cost.

The progression from CNNs through ViT to Swin illustrates a consistent tension. Richer spatial context requires longer-range interactions, and longer-range interactions are computationally expensive. CNNs are cheap but local. ViT is global but quadratic. Swin is a practical compromise but still grows with resolution and fixes its context window at design time. For a recognition backbone that needs to process high-resolution images with globally distributed symptoms across a leaf surface, this tension is not merely theoretical. A backbone that cannot efficiently model long-range spatial dependencies will systematically fail to capture the whole-leaf patterns that distinguish a moderately infected leaf with scattered lesions from a severely infected one with confluent lesion areas and secondary chlorosis. This is the architectural context that motivates looking beyond attention-based backbones toward alternatives with linear complexity and content-adaptive long-range modelling.

2.1.4 State Space Models and Visual State Space Models (VMamba)

The alternative to attention-based long-range modelling comes from control theory and signal processing. Continuous-time linear state space models have a long history in classical control, but their application to deep sequence modelling was formalised by Gu et al. (2022) with the Structured State Space Model (S4). The core idea is to model a sequence through a continuous-time dynamical system:

$$x'(t) = \mathbf{A}x(t) + \mathbf{B}u(t) \tag{2.1}$$

$$y(t) = \mathbf{C}x(t) + \mathbf{D}u(t) \tag{2.2}$$

where $u(t)$ is the input signal, $x(t)$ is a hidden state that summarises the history of the input, and $y(t)$ is the output. The state matrix $\mathbf{A}$ is parameterised using the HiPPO framework (Gu et al., 2022), which initialises it to project the input history onto an optimal basis of orthogonal polynomials. This gives S4 long-range memory without the gradient degradation that cripples recurrent networks like LSTMs. Critically, S4

can be computed as a global convolutional kernel during training, enabling $O(N \log N)$ FFT-based parallelism, while at inference it runs as a recurrent network in $O(N)$ time.

S4's limitation is that its parameters $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ are fixed regardless of the input content. It applies the same transformation to every sequence, making it a strong memory mechanism but a weak content-based reasoning tool. Gu and Dao (2024) corrected this with Mamba, which introduces a selective scan mechanism (S6) where the parameters B, C, and the discretisation timescale Δ are computed as learned linear projections of the current input:

$$B = s_B(x), \quad C = s_C(x), \quad \Delta = \text{softplus}(s_\Delta(x)) \tag{2.3}$$

A small Δ causes the model to strongly incorporate the current input into the hidden state. A large Δ causes it to decay the current input and retain prior state. This input-dependent gating gives Mamba the ability to selectively propagate or suppress information based on content, an ability that S4 completely lacked, while preserving $O(N)$ complexity through a hardware-aware parallel scan. On language modelling benchmarks, Mamba achieved Transformer-level perplexity at linear cost and was the first SSM to match Transformers on information-dense data.

S4 and Mamba were both designed for one-dimensional sequences such as text, audio, and time series. Their relevance to the vision backbone question arises through the challenge of adapting the selective scan to two-dimensional image data. VMamba (Liu et al., 2024) brings the Mamba selective scan into the image domain through a targeted solution to the dimensionality mismatch. When a 2D image is naively flattened into a 1D sequence for processing by S6, artificial causal dependencies arise. Pixels near the end of the flattened sequence receive context from earlier pixels but cannot contribute context back, breaking the omnidirectional spatial symmetry that image understanding requires. A leaf lesion at the bottom-right of an image should inform the representation of a related symptom at the top-left, regardless of which position comes first in a scanning order.

VMamba solves this with the **2D Selective Scan (SS2D)** module, which pro-

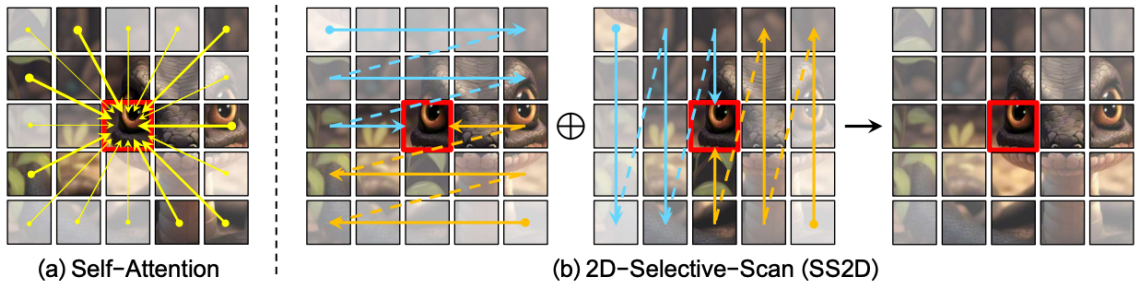


Figure 2.3: VMamba's SS2D module: four-directional selective scan for omnidirectional spatial context.

cesses the spatial feature map independently along four traversal directions: top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right. The four resulting hidden-state sequences are merged into a single spatial output, so that every position receives context from the full spatial extent of the feature map through at least one scanning direction. The result approximates omnidirectional spatial attention at $O(N)$ cost rather than $O(N^2)$. Figure 2.3 illustrates the four-direction scan. VMamba organises its VSS blocks (Visual State Space blocks based on SS2D) into a four-stage hierarchical architecture with patch-merging downsamplers between stages, directly mirroring Swin's hierarchical design and producing the same multi-scale feature pyramid.

The choice of VMamba-Small for the recognition backbone of DCSC in this thesis is motivated by three considerations. First, VMamba's SS2D mechanism operates at $O(N)$ complexity, whereas Swin's shifted-window attention is $O(N \cdot M^2)$ and requires the window size M to be fixed at design time. For plant disease images where discriminative features can span the entire leaf surface, a backbone that avoids hard-coded spatial context limits is preferable. Second, VMamba-Small matches or exceeds Swin Transformer-Small on ImageNet-1K classification, COCO object detection, and ADE20K semantic segmentation, so any accuracy gain observed in DCSC over the most directly competing baseline (TSTC, which uses a Swin backbone) can be attributed to the conditioning mechanism rather than to a stronger backbone. If DCSC used a stronger backbone and TSTC a weaker one, the comparison would be confounded. Third, the content-sensitive per-token gating in VMamba's selective scan

means that the backbone naturally suppresses uninformative background regions and amplifies lesion regions, a property that aligns with the spatially non-local nature of severity assessment across a whole-leaf image.

2.1.5 What VMamba Lacks: Cross-Task Conditioning

Despite its architectural advantages, VMamba shares a structural limitation with every other visual backbone reviewed in this section: it produces a single unified representation $\mathbf{h} \in \mathbb{R}^D$ from which independent task heads branch. There is no mechanism in VMamba by which the computation of one task's prediction can modify the feature space used by another. In a multitask setting where the optimal decision boundary for one task depends on the outcome of a co-predicted task, this independence is not a harmless simplification. In the plant disease multitask setting specifically, the severity head must decide whether a leaf shows general or serious infection using the same feature vector that the disease head uses for its own prediction, with no mechanism to condition that severity decision on what the disease head found. The result is a single disease-agnostic severity boundary, and the systematic confusion between general and serious severity that this boundary produces is documented in the TSTC and LDI-NET confusion matrices in Section 2.3. This gap is the primary architectural motivation for the first contribution of this thesis. The next section reviews the multitask learning and feature conditioning literature to identify the appropriate mechanism for filling it.

2.2 Multitask Learning, Feature Conditioning, and Ordinal Supervision

The gap identified in Section 2.1.5 is a conditioning problem. The severity feature space needs to be transformed differently for each disease class, and the transformed features must be supervised in a way that respects the ordinal structure of the severity label space. This section reviews three threads of literature that together provide the

tools for both requirements: multitask learning frameworks that motivate cross-task feature interaction, FiLM-based conditioning mechanisms that implement the discrete-label-conditioned transformation, and ordinal regression methods that supervise the resulting predictions in a rank-consistent way.

2.2.1 Multitask Learning in Computer Vision

Multitask learning (MTL) in visual recognition was formalised in early work by (Caruana, 1997) and has been extensively revisited in the deep learning era (Crawshaw, 2020). The dominant paradigm is hard-parameter-sharing: a single backbone encoder processes the input image and feeds task-specific prediction heads from a common representation. Hard sharing is computationally efficient and provides implicit regularisation, because the shared backbone is forced to extract features that are simultaneously useful for all tasks. Its central assumption, however, is that all tasks are best served by the same representation, which fails when tasks have different or even conflicting feature requirements.

Soft-parameter-sharing methods (Misra et al., 2016) relax this assumption by allowing separate per-task networks while penalising their parameters for diverging too far, giving each task some representational independence without complete separation. Cross-stitch networks (Misra et al., 2016) go further by learning task-specific linear combinations of features from multiple task branches, which is a step toward cross-task feature interaction. Task-grouping methods (Standley et al., 2020) use gradient interference analysis to decide which tasks should share a backbone at all. None of these approaches model an explicit directional conditional dependency between tasks, specifically the scenario where one task's decision boundary should be dynamically transformed based on the discrete output of a co-predicted task. For the disease-severity problem, what is needed is not a shared representation or a linear feature combination but a mechanism that applies a different feature transformation for each value of the disease label. This is a structured conditioning problem.

2.2.2 Feature-wise Linear Modulation (FiLM)

With the conditioning problem defined precisely as a discrete-label-conditioned feature transformation, the choice of mechanism comes down to comparing the available options. Cross-attention allows severity features to attend over a learned disease label embedding, but cross-attention requires the conditioning signal to be a sequence of vectors. When the conditioning signal is a single scalar index, as in DCSC where the predicted disease class is a single label, cross-attention reduces to a single global weighted sum, which is mathematically equivalent to a linear projection and offers no advantage over simpler alternatives. Concatenation conditioning appends the disease embedding to the feature vector before the severity head, but it entangles the conditioning signal and the features in a way that requires the head to implicitly learn to disentangle them, with no architectural guarantee that this disentanglement occurs. Hypernetwork conditioning uses a secondary network to generate task-specific weights for the severity head, which is the most flexible option but also the most parameter-expensive and the most difficult to train stably.

Feature-wise Linear Modulation (FiLM), introduced by Perez et al. (2018), provides the right balance. FiLM applies a learned feature-wise affine transformation to intermediate network activations:

$$\tilde{\mathbf{f}} = \boldsymbol{\gamma}(c) \odot \mathbf{f} + \boldsymbol{\beta}(c) \tag{2.4}$$

where $\mathbf{f}$ is the feature map to be conditioned, c is the conditioning signal (in DCSC, the predicted disease class), and $\boldsymbol{\gamma}(c)$ and $\boldsymbol{\beta}(c)$ are scale and shift vectors generated by a lightweight conditioning network: a single embedding lookup followed by a two-layer MLP. The operation is interpretable: $\gamma_i > 1$ amplifies dimension i of the severity feature space for disease c, while β_i shifts its baseline activation. Different disease classes produce different $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ pairs, effectively creating disease-specific feature subspaces within the shared backbone representation, without modifying the backbone itself. Figure 2.4 shows the FiLM conditioning block.

FiLM was originally demonstrated on the CLEVR visual question answering bench-

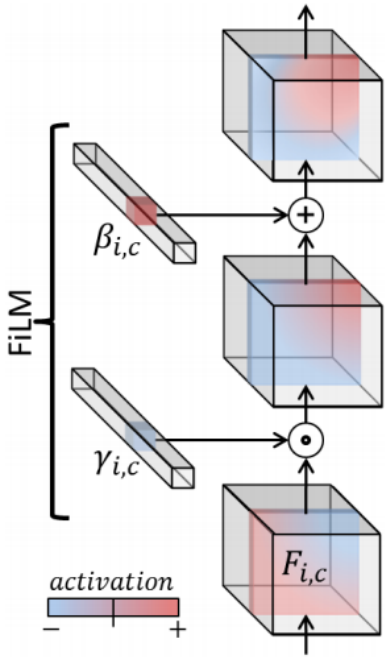


Figure 2.4: Feature-wise Linear Modulation (FiLM) block.

mark (Perez et al., 2018), where conditioning visual features on natural language questions produced state-of-the-art results without any architectural change to the visual encoder, confirming that affine feature modulation is sufficient to implement complex cross-modal conditioning. Durasov et al. (2022) later demonstrated FiLM's versatility across probabilistic deep learning settings. To the best of our knowledge, this thesis is the first to apply FiLM in a cross-task setting where the conditioning signal is a co-predicted categorical label rather than an external query.

2.2.3 Conditional Multitask Learning and Prior Cross-Task Work

The conceptual precedent for conditioning one task's prediction on another's output was established for plant disease analysis by Lee et al. (2021), who proposed Conditional Multi-Task Learning (CMTL). CMTL conditions disease classification on the predicted species identity, demonstrating that encoding the hierarchical rela-

tionship between tasks as an explicit conditional dependency, rather than relying on shared-backbone implicit co-learning, improves performance on both tasks. DCSC extends this conditioning logic one level further: from species-to-disease conditioning (as in CMTL) to disease-to-severity conditioning. This is the more agronomically consequential direction, because general–serious severity confusion can affect the choice and timing of intervention.

Tripathi et al. (2024) proposed a multitask vision transformer for simultaneous disease localisation and classification, confirming that explicit task interaction outperforms independent heads on joint plant disease tasks. The localisation-classification coupling in that work connects spatial and semantic tasks rather than two semantic tasks with a conditional dependency, but it reinforces the broader argument that cross-task interaction improves multitask plant disease models.

2.2.4 Ordinal Regression and Consistent Rank Logits

Severity is not a nominal categorical variable. The label space healthy to general to serious is ordered. The visual transition from healthy to general involves the first appearance of lesion symptoms, while the transition from general to serious involves secondary effects such as chlorosis and tissue necrosis. Adjacent classes are more visually similar than distant classes, and this ordering carries real agronomic meaning. Confusing general with serious leads to over-treatment, applying fungicide to a field that does not need it. Confusing serious with general leads to under-treatment, failing to intervene in a field that requires immediate action. Standard cross-entropy loss treats all misclassifications as equally bad, assigning the same penalty to a serious-versus-healthy error (an ordinal distance of 2) as to a general-versus-healthy error (an ordinal distance of 1). This is incorrect for severity grading and motivates the use of an ordinal regression framework in the DCSC severity head.

Cao et al. (2020) proposed CORAL (Consistent Rank Logits), which decomposes ordinal prediction into a set of binary cumulative subtasks. For severity labels $s \in$

$\{0, 1, 2\}$ corresponding to healthy, general, and serious, CORAL estimates:

$$P(\hat{y}_s > k) = \sigma(\mathbf{w}^\top \tilde{\mathbf{f}}_{\text{sev}} + b_k), \quad k \in \{0, 1\} \tag{2.5}$$

using a shared weight vector $\mathbf{w}$ and rank-specific bias scalars b_k. The shared weight constraint guarantees that the cumulative probabilities satisfy $P(\hat{y}_s > 0) \geq P(\hat{y}_s > 1)$ by construction. This prevents rank-inconsistent cumulative outputs and encourages ordinal coherence; it does not, by itself, eliminate ordinary class misclassification between adjacent severity grades.

Shi et al. (2021) proposed CORN as a relaxation that removes the shared-weight constraint and preserves rank consistency through conditional probability decomposition. CORN is more expressive but also more parameter-heavy and loses the hard architectural guarantee of rank-monotonicity. For the DCSC severity head, CORAL is preferred because the guarantee rather than the flexibility is what matters. The target is to eliminate a specific confusable class pair (general versus serious), and CORAL's shared-weight construction directly addresses this. Ordinal regression has also been applied effectively in related severity grading domains. Gao et al. (2024) applied a CORAL-based framework to cardiovascular disease risk stratification from retinal fundus images, showing that ordinal supervision provides consistent calibration advantages over cross-entropy even when backbone performance is held constant. Diabetic retinopathy grading (Bellaj et al., 2026) has shown the same pattern across multiple backbone architectures. To the best of our knowledge, ordinal regression had not been applied to plant disease severity grading prior to this thesis.

2.3 Plant Disease Classification and Severity Estimation

This section situates the thesis within the broader plant disease vision literature. The single-task classification subsection reviews the backbone trajectory inherited from the wider literature and explains why flat compound labels limit prior approaches.

The multitask subsection identifies the specific prior models that DCSC is compared against and the shared architectural assumption that none of them has yet broken.

2.3.1 Single-Task Plant Disease Classification

Automated plant disease classification has closely tracked the evolution of visual backbones reviewed in Section 2.1. The earliest approaches used hand-crafted features such as colour histograms, local binary patterns, and Gabor texture descriptors combined with SVM classifiers (Gulhane and Gurjar, 2011; Dutot et al., 2013; Sabrol and Satish, 2016; Kurmi et al., 2022). These worked reasonably well in controlled laboratory imaging conditions but degraded under field variability. Inconsistent illumination, background clutter, leaf overlap, and the large within-class visual variation of field-collected disease samples all caused performance to drop sharply (Brahimi et al., 2018; Picon et al., 2019).

The PlantVillage dataset (Hughes and Salathé, 2015), containing over 54,000 lab-photographed leaf images across 26 diseases, established the first large-scale benchmark and enabled CNN-based approaches to demonstrate near-human performance in the controlled setting. Deep CNNs pre-trained on ImageNet and fine-tuned on PlantVillage achieved classification accuracies above 99% in the lab (Mohanty et al., 2016), though performance dropped substantially when the same models were applied to field-collected images (Brahimi et al., 2018). Subsequent work improved field robustness through data augmentation, domain adaptation, and architectures with larger effective receptive fields (Ferentinos, 2018; Chen et al., 2020; Yang et al., 2020).

A structural limitation shared by all single-task CNN classifiers is the flat label formulation: a single softmax output over compound labels that conflate species, disease, and severity into one category index. A model trained with this formulation cannot exploit the natural hierarchical structure of the label space, because the relationships between categories such as apple and tomato sharing no diseases, or rust and early blight requiring disease-specific severity criteria, are invisible to the loss function. This limitation directly motivates the multitask formulations reviewed next.

Vision Transformers subsequently surpassed CNN baselines on multiple agricul-

tural benchmarks. The Swin Transformer proved particularly well-suited to leaf disease images, where discriminative features span multiple spatial scales, from fine-grained lesion textures at the microscopic level to coarse lesion distributions at the whole-leaf level. Tai et al. (2023) achieved state-of-the-art results on cassava leaf disease classification with a lightweight vision transformer, and Wang et al. (2022b) demonstrated strong performance on cucumber leaf diseases under small-sample conditions. Despite these improvements, all these models remain single-task classifiers that carry the flat-label limitation described above.

2.3.2 Multitask Disease-Severity Models

The limitation of flat compound labels motivated the development of multitask formulations for joint plant disease and severity prediction. The relevant prior models reviewed below share a structural assumption that none of them has yet broken: all impose a single shared severity decision boundary across all disease classes.

Liang et al. (2019) proposed PD^{2}SE-Net, the first large-scale multitask model for joint disease and severity prediction, built on ShuffleNet-v2. PD^{2}SE-Net uses independent prediction heads from a shared backbone with no mechanism for disease-to-severity conditioning. Wang et al. (2022a) extended this with the Dual-Stream Hierarchical Bilinear Pooling model (DHBP), which uses compact bilinear pooling to model feature interactions between disease and severity feature streams. DHBP is a step toward inter-task feature interaction, but the bilinear pooling operation is symmetric and global: it models a joint feature distribution rather than conditioning the severity feature space on the discrete disease prediction, which is what the asymmetric conditional dependency requires. Keceli et al. (2022) proposed a multi-branch system for joint crop and disease classification, confirming the utility of task-specific branches but again without cross-task conditioning.

The most directly relevant prior work is the Triple-branch Swin Transformer Classification (TSTC) network of Yang et al. (2023). TSTC uses a Swin Transformer backbone with compact bilinear pooling and a five-loss deep supervision module for simultaneous disease-type and severity classification. On AI Challenger 2018 with

an 80/20 random split, TSTC achieves 99.00% disease accuracy and 88.73% severity accuracy, which represents the prior state of the art. In the fixed-split reproduction analysed in Chapter 3, the TSTC severity confusion matrix contains 335 instances of *serious* predicted as *general* and 187 instances of *general* predicted as *serious*. The pronounced serious-to-general error is consistent with a disease-agnostic severity boundary. TSTC's near-ceiling disease accuracy suggests that this severity confusion is not simply a consequence of weak disease recognition; the architecture provides no mechanism for adapting the severity boundary to the predicted disease.

Yang et al. (2024b) proposed LDI-NET, a single-branch multi-label CNN-Transformer framework that simultaneously predicts species, disease, and severity. LDI-NET reports a *general*-severity accuracy of only 81.49%, with overall severity accuracy of 88.55%, confirming that the general-serious confusion persists even with a stronger and more complex architecture. Both TSTC and LDI-NET share the same structural assumption: a single disease-agnostic severity decision boundary. DCSC is the first model to reject this assumption explicitly, and the combination of FiLM-based cross-task conditioning with CORAL ordinal supervision proposed in Chapter 3 is designed to address both the disease-conditioned boundary problem and the ordinal consistency problem simultaneously.

2.4 Generative Modelling for Data Augmentation

The second contribution of this thesis addresses the generative side of plant disease analysis. Field-collected agricultural datasets are often highly class-imbalanced, with early-stage disease cases being particularly scarce. This scarcity limits both model training and evaluation, since the most agronomically important cases are also the least represented in the data. The central question, therefore, is how to synthesise early-stage disease imagery that is not only visually realistic but also morphologically faithful to expert descriptions of disease symptoms.

KGDC-Diffusion is proposed to address this problem. While DCSC focuses on recognition by conditioning severity estimation on disease identity, KGDC-Diffusion

focuses on data generation by conditioning synthesis on retrieved morphological knowledge. This section reviews generative modelling approaches for agricultural data augmentation, identifies the conditioning failures that motivate the KGDC-Diffusion design, and justifies the use of retrieval-augmented conditioning.

2.4.1 Generative Adversarial Networks

Generative Adversarial Networks (Goodfellow et al., 2014) established the first practical framework for high-fidelity image synthesis. A GAN trains a generator network G to produce images that fool a discriminator network D, with D trained simultaneously to distinguish generated images from real ones. The adversarial objective pushes G toward producing images that lie within the real image distribution. Conditional GANs (Mirza and Osindero, 2014) extended this framework by conditioning both G and D on a class label, enabling category-specific generation.

For agricultural augmentation specifically, GAN-based approaches have shown consistent but modest gains. Lu et al. (2022) survey the field comprehensively, Espejo-Garcia et al. (2021) combined DCGAN with transfer learning for weed image generation, and Abbas et al. (2021) used a Conditional GAN to synthesise tomato disease images, reporting 1 to 4% downstream classification accuracy improvements. These results confirm that generative augmentation is a viable strategy for agricultural vision, but they do not address the fundamental limitation of GANs in the low-data regime: mode collapse. When training data for a target class is scarce, the GAN generator converges to a small number of high-confidence output modes that satisfy the discriminator, rather than exploring the full diversity of the target distribution (Creswell et al., 2018). For early-stage disease images, the defining visual features are precisely the subtle, diverse signs that appear before infection consolidates into a recognisable canonical pattern. Mode collapse eliminates exactly this diversity. This is confirmed quantitatively in Chapter 4: StyleGAN3 achieves high Precision (0.82) but Recall of only 0.09 on the SoySeverityDataset, meaning it generates images that look individually plausible but covers only a small fraction of the real data distribution.

2.4.2 Latent Diffusion Models and Spatial Conditioning

The mode collapse and training instability problems of GANs were substantially resolved by Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020). A DDPM defines a forward process that gradually adds Gaussian noise to a real image over T timesteps and trains a neural network to reverse this process by predicting and subtracting the noise at each step. Because the training objective is a simple mean-squared error on noise prediction rather than an adversarial game, the training dynamics are stable and the learned model covers the full data distribution rather than collapsing to a subset of modes. Figure 2.5 illustrates the forward and reverse processes. Dhariwal and Nichol (2021) showed that classifier-guided diffusion models surpass GANs on FID, Precision, and Recall across all standard benchmarks, with the Recall gains being as large as the FID gains, confirming that diffusion models genuinely improve distributional coverage rather than just per-sample quality.

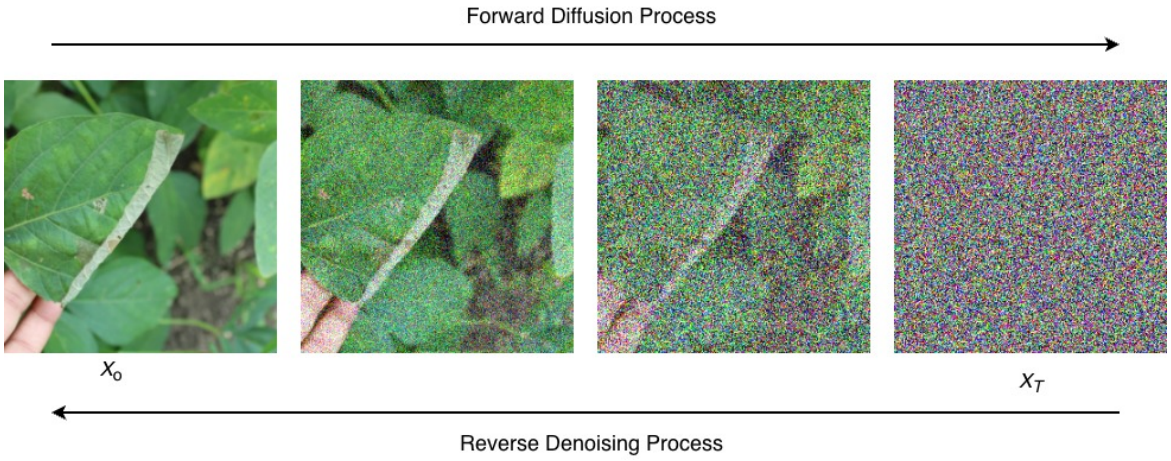


Figure 2.5: DDPM forward and reverse diffusion processes.

Latent Diffusion Models (LDMs) (Rombach et al., 2022) improve the computational feasibility of diffusion by shifting the process into the compressed latent space of a pre-trained variational autoencoder. With a spatial compression factor of $f = 8$, the number of spatial dimensions is reduced by $f^2 = 64$, making fine-tuning feasible on a single GPU rather than requiring multi-node infrastructure. The denoising U-Net within an LDM is conditioned on a text prompt through CLIP (Radford et al., 2021)

embeddings injected via cross-attention layers:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathcal{E}(\mathbf{x}),\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \tau_\theta(\mathbf{y}))\|^2\right] \tag{2.6}$$

where $\tau_\theta(\mathbf{y})$ encodes the conditioning signal. ControlNet (Zhang et al., 2023) extends LDMs with additional spatial conditioning inputs including segmentation masks, depth maps, and edge maps, through zero-initialised trainable adapter layers appended to the frozen U-Net encoder. The zero-initialisation ensures that at the start of ControlNet training, the model's output is identical to the base LDM, preventing catastrophic forgetting of the pre-trained weights. ControlNet is directly relevant to KGDC-Diffusion because it provides a mechanism for constraining the spatial placement of synthesised lesions on the leaf surface, not merely their semantic category. Figure 2.6 shows the ControlNet adapter design.

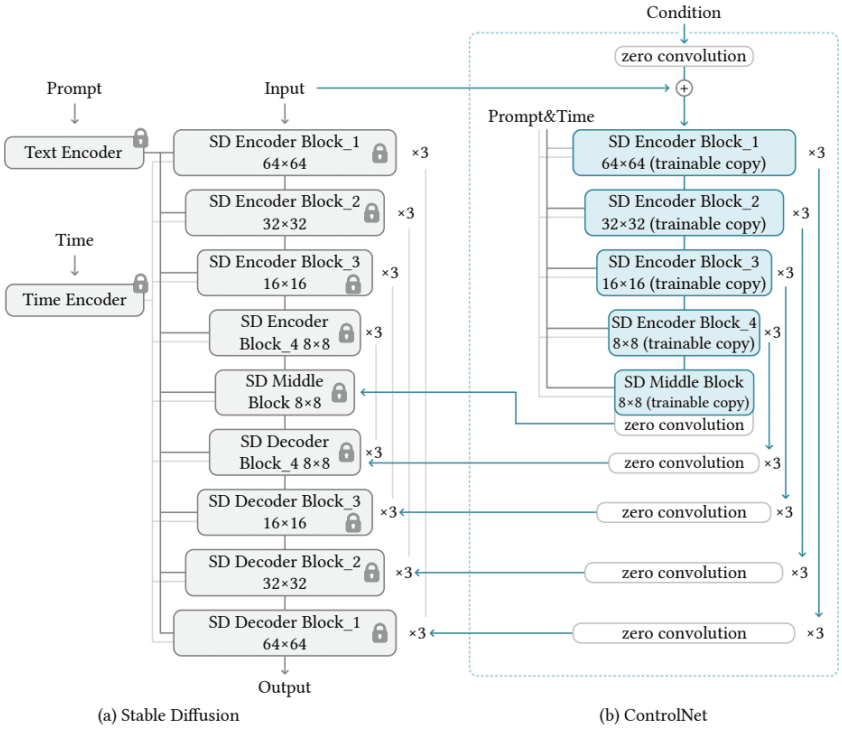


Figure 2.6: ControlNet adapter architecture for spatial conditioning of Latent Diffusion Models.

Chen et al. (2024) conducted the first systematic evaluation of diffusion models

for weed image generation, demonstrating that diffusion-augmented training sets outperform GAN-augmented sets across four downstream classifiers. This establishes diffusion models as the appropriate generative framework for agricultural augmentation and sets the stage for the remaining question: what conditioning signal should be used?

2.4.3 The Conditioning Insufficiency Problem

Despite the clear advantages of diffusion models over GANs in terms of sample diversity and training stability, all existing agricultural diffusion models condition generation on discrete class labels or generic text strings. This is a fundamentally impoverished conditioning signal. Consider what happens when a latent diffusion model trained on soybean disease images receives the class label *early blight stage 1* as its conditioning input. The label tells the model nothing specific about what early blight looks like at stage 1: nothing about the water-soaked brown margins, the concentric target rings at the lesion centre, the pale chlorotic halo surrounding each lesion, or the typical lesion diameter at this stage. The model is supposed to learn these visual properties from the training images associated with that label. But the early-stage classes are precisely the classes with the fewest training images, because early infections are transient and difficult to photograph at exactly the right moment in the field. The model therefore has both the least guidance from its conditioning signal and the least training data from which to learn compensating priors. The result is a **conditioning insufficiency problem**: generated images look like plausible plant leaves and may even look diseased, but they fail to reproduce the subtle, specific visual signatures that characterise early-stage infection and differentiate it from late-stage infection of a different disease.

This problem is not unique to plant pathology. In medical imaging, synthesis of rare or early-stage findings faces the same challenge. A label such as *mild diabetic retinopathy* does not specify what mild retinopathy looks like in terms of microaneurysm count, haemorrhage distribution, or vessel calibre changes, and early-stage retinopathy images are exactly the scarce class. The conditioning insufficiency prob-

lem is therefore a broadly relevant open problem in conditional generative modelling for specialised scientific domains.

2.4.4 Retrieval-Augmented Generation as a Conditioning Strategy

Given that the problem is an impoverished conditioning signal, the natural question is where a richer signal should come from. Three alternative strategies are conceivable. The first is to fine-tune the CLIP text encoder on domain-specific plant pathology text, giving it better statistical coverage of domain vocabulary. The second is to use a vision-language model to generate image-specific descriptive captions at training time, which are then used as conditioning prompts. The third is to retrieve domain-specific knowledge from a curated expert corpus at inference time and use the retrieved text as the conditioning prompt. The first option requires substantial compute and data to fine-tune a large vision-language model. The second option requires a VLM that can accurately describe early-stage disease morphology from images, but early-stage images are scarce and are unlikely to be well-represented in the VLM's training distribution, so the generated descriptions will be generic. The third option, retrieval, does not require retraining any large model. The knowledge base is authored by a domain expert, validated for accuracy, and used at inference time without any modification to the base diffusion model.

Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) was developed in natural language processing to ground language model outputs in dynamically retrieved factual context, mitigating hallucination on knowledge-intensive tasks. The core mechanism encodes a curated knowledge base into dense vector embeddings, then at query time retrieves the most semantically relevant documents by cosine similarity and injects their content into the model's prompt. This allows the model to produce outputs grounded in specific, expert-verified facts without retraining. The use of retrieved botanical morphological descriptors to condition a latent diffusion model—replacing bare class labels with expert-authored visual specifications retrieved for each query—is

the specific contribution of the KGDC-Diffusion framework. To the best of our knowledge, this represents the first application of retrieval-augmented conditioning to a generative vision model in the botanical domain.

Chapter 3

Visual State Space Model for Disease-Conditioned Severity Identification

3.1 Introduction

Severity-aware disease identification requires more than recognising the disease present in an image. It also requires estimating how far the disease has progressed, and this estimation is inherently dependent on the disease identity itself. The same visible extent of damage may indicate different severity levels for different pathogens, because each disease produces distinct lesion structures, colour changes, spatial spread patterns, and tissue responses. This chapter introduces Disease-Conditioned Severity Calibration (DCSC), a VMamba-based multitask framework that makes this dependency explicit by conditioning the severity prediction pathway on the predicted disease identity.

Chapter 2 showed that this dependency is absent from prior plant disease multitask models. Existing approaches treat severity as a disease-agnostic categorical task and apply a single shared severity decision boundary across all disease classes. The result is the systematic general-to-serious confusion pattern documented in the TSTC (Yang et al., 2023) and LDI-NET (Yang et al., 2024b) confusion matrices, with hundreds of severity errors in each direction across thousands of test images. DCSC addresses this

limitation directly by replacing a disease-agnostic severity pathway with a disease-conditioned one.

The core diagnostic insight behind the design is one that practising plant pathologists apply routinely. When a pathologist grades disease severity in the field, they do not use a universal visual ruler. They first identify the disease, then apply disease-specific criteria for what general and serious severity look like for that particular pathogen. A scab infection occupying twenty percent of the leaf surface may be classified as serious because the lesions are deep and necrotic, while a rust infection at the same coverage may be classified as general because the pustules are shallow and uniformly distributed. The visual manifestation of severity is therefore conditional on disease identity. This observation is intuitive to a domain expert but is not explicitly encoded in existing multitask plant disease architectures.

DCSC encodes this prior architecturally. The framework uses the VMamba-Small backbone justified in Section 2.1 of Chapter 2 to produce a shared visual representation, applies a disease classification head, and then modulates the severity feature space through Feature-wise Linear Modulation (FiLM) conditioned on the predicted disease identity. The modulated severity features feed a two-head severity stack, a CORAL ordinal regression head that encourages rank-consistent cumulative predictions and a disease-conditioned prototype head that imposes ordinal geometry in feature space. The class probabilities are then fused, with an Earth Mover's Distance loss providing distributional smoothing. On AI Challenger 2018 under a five-seed 70/15/15 fixed-split protocol, DCSC achieves 98.51% disease accuracy and 92.30% severity accuracy with a soft severity MAE of 0.0909. On the Rice Leaf Severity dataset, where disease accuracy reaches the ceiling for all evaluated methods, DCSC achieves a soft severity MAE of 0.031, a reduction of 79% over the VMamba baseline.

3.2 Proposed Methodology

The DCSC framework comprises a shared visual backbone, a disease classification head, a FiLM-based disease-conditioned severity modulation block, and a two-head

severity stack combining a CORAL ordinal regression head and a disease-conditioned prototype head, trained alongside an EMD distributional smoothing loss. The architecture is shown in Figure 3.1. This section first formalises the joint prediction problem, then describes the overall architecture, and finally specifies each of the innovations (FiLM modulation, CORAL ordinal regression, disease-conditioned severity prototypes, and EMD smoothing) along with the combined training objective.

Let $\mathcal{X}$ denote the space of input leaf images, $\mathcal{Y}_d = \{0, 1, \ldots, K_d - 1\}$ the discrete space of K_d disease classes, and $\mathcal{Y}_s = \{0, 1, 2\}$ the ordinal severity space with 0 corresponding to healthy, 1 to general, and 2 to serious. The number of disease classes K_d is dataset-specific: $K_d = 20$ for AI Challenger 2018 and $K_d = 4$ for the Rice Leaf Severity dataset. The joint multitask prediction problem is to learn a function $f : \mathcal{X} \to \mathcal{Y}_d \times \mathcal{Y}_s$. In existing multitask formulations such as TSTC and LDI-NET, f is decomposed into a shared backbone $\phi : \mathcal{X} \to \mathbb{R}^D$ that produces a D-dimensional representation $\mathbf{h}$, followed by two independent prediction heads $g_d : \mathbb{R}^D \to \mathcal{Y}_d$ and $g_s : \mathbb{R}^D \to \mathcal{Y}_s$. The independence assumption is that $g_s(\mathbf{h})$ does not depend on $g_d(\mathbf{h})$, so the severity head applies a single decision boundary across all disease classes. The DCSC formulation rejects this independence assumption. The severity head is reformulated as $g_s(\mathbf{h}, \hat{y}_d)$ where $\hat{y}_d = g_d(\mathbf{h})$ is the predicted disease identity, and the conditioning is implemented as a feature transformation of the severity branch rather than as an additional input concatenation.

3.2.1 Architecture Overview

The backbone ϕ is VMamba-Small (Liu et al., 2024), which produces a $D = 768$ dimensional feature representation $\mathbf{h} \in \mathbb{R}^D$ after global average pooling of the final stage output. To summarise the justification developed in Section 2.1: VMamba's 2D Selective Scan operates at linear computational cost with respect to spatial resolution, it matches Swin Transformer-Small on standard recognition benchmarks so backbone strength is held constant against the most directly competing prior work, and its content-sensitive selective gating aligns with the spatially non-local nature of severity assessment across a whole-leaf image. The backbone is initialised from the ImageNet-

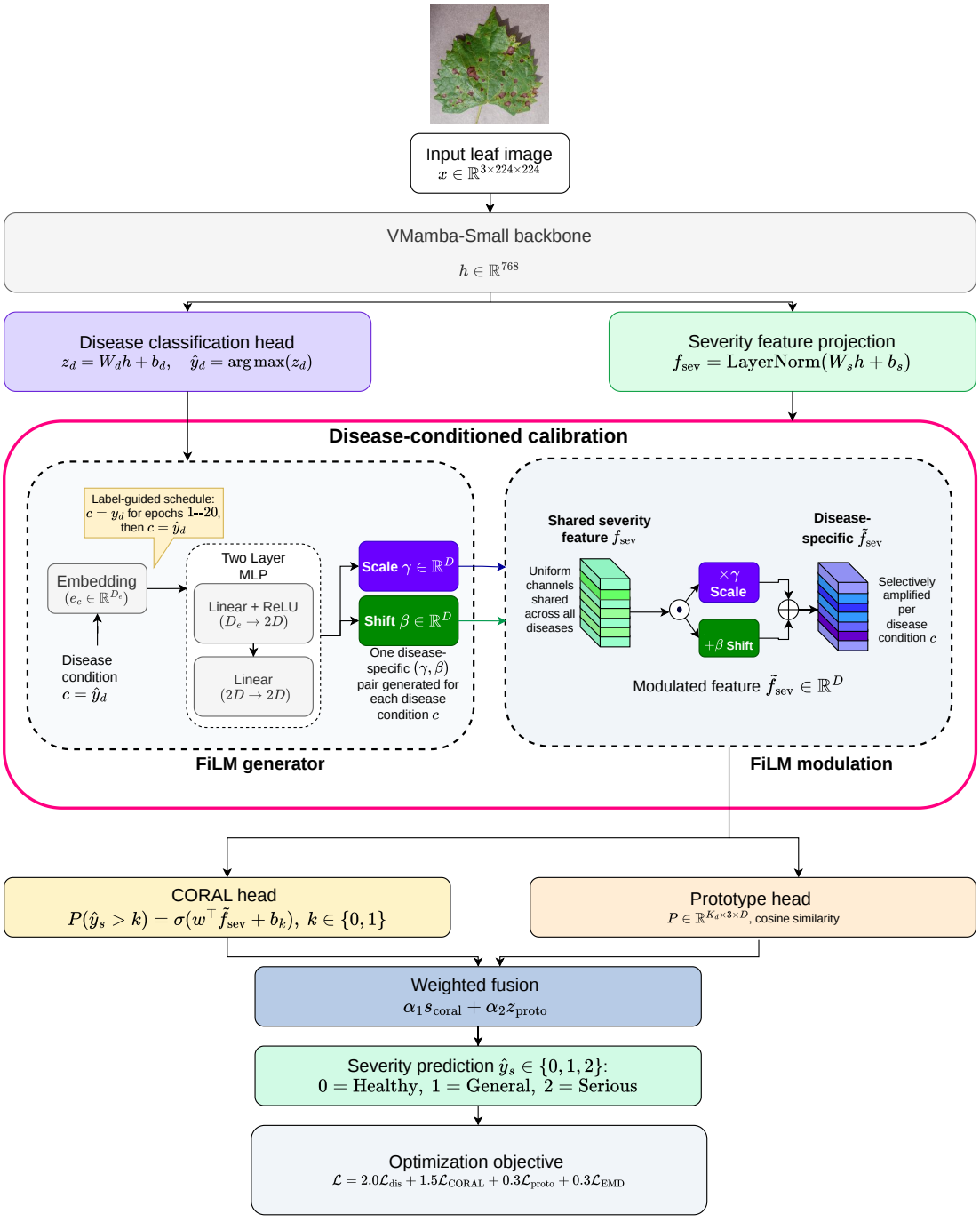


Figure 3.1: DCSC architecture. The shared VMamba-Small backbone feeds a disease head and a severity projection. The predicted disease identity drives a FiLM generator that modulates the severity feature space. CORAL cumulative outputs are converted to a three-class distribution and fused with the disease-conditioned prototype distribution. Training uses disease cross-entropy, CORAL, prototype, and Earth Mover's Distance objectives.

1K pre-trained checkpoint and fine-tuned end-to-end as part of the DCSC training procedure. Both the disease and severity heads draw from the same pooled feature $\mathbf{h}$, ensuring that the backbone learns a single representation that supports both tasks rather than encoding task-specific features in disjoint regions.

The disease head is a single linear layer applied to the pooled backbone feature:

$$\mathbf{z}_d = \mathbf{W}_d\mathbf{h} + \mathbf{b}_d, \quad \mathbf{W}_d \in \mathbb{R}^{K_d \times D}, \quad \mathbf{b}_d \in \mathbb{R}^{K_d}. \tag{3.1}$$

The predicted disease identity is obtained by the standard argmax decision $\hat{y}_d = \arg\max_k \mathbf{z}_d[k]$. The simplicity of this head is intentional. Disease classification on the benchmarks considered here is close to the ceiling for any reasonable backbone (TSTC reports 99% and DCSC reports 98.51%), so a richer disease head would yield no measurable improvement. The architectural innovation of DCSC is therefore concentrated on the severity pathway where the structural gap lies.

3.2.2 FiLM-Based Disease-Conditioned Severity Modulation

The pooled backbone feature is projected into a severity-specific feature space through a layer-normalised linear transformation:

$$\mathbf{f}_{\text{sev}} = \text{LayerNorm}(\mathbf{W}_s\mathbf{h} + \mathbf{b}_s), \quad \mathbf{W}_s \in \mathbb{R}^{D \times D}. \tag{3.2}$$

The disease conditioning signal is the predicted disease identity c. During the first 20 training epochs, c is set to the ground-truth disease label y_d to prevent noise from an untrained disease head from disrupting the severity modulation pathway. After epoch 20 the conditioning signal switches to the predicted disease identity $\hat{y}_d$, matching the inference setting. This curriculum design is consistent with standard practice in conditional training where the conditioning signal is itself a learned prediction.

The conditioning signal is embedded and passed through a two-layer multilayer

perceptron that produces the FiLM scale and shift parameters:

$$\mathbf{e}_c = \text{Embedding}(c) \in \mathbb{R}^{D_e}, \quad [\boldsymbol{\gamma}, \boldsymbol{\beta}] = \text{MLP}_{\text{FiLM}}(\mathbf{e}_c), \quad \boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^D. \tag{3.3}$$

The modulated severity feature is the element-wise affine transformation:

$$\tilde{\mathbf{f}}_{\text{sev}} = \boldsymbol{\gamma} \odot \mathbf{f}_{\text{sev}} + \boldsymbol{\beta}, \tag{3.4}$$

where $\odot$ denotes element-wise multiplication. The interpretation is direct. For each disease class c, the FiLM generator produces a specific pair $(\boldsymbol{\gamma}_c, \boldsymbol{\beta}_c)$ that scales and shifts each dimension of the severity feature space independently. Different diseases produce different modulation parameters, so the effective severity feature space differs across diseases. A general lesion in rust and a general lesion in early blight, after passing through their respective FiLM transformations, occupy different regions of the modulated feature space, and the downstream severity decision boundary can be calibrated for each disease individually. The justification for choosing FiLM over alternatives such as cross-attention, concatenation conditioning, or hypernetwork conditioning was developed in Section 2.2.2 of Chapter 2.

3.2.3 CORAL Ordinal Severity Head

The modulated severity feature feeds a CORAL (Cao et al., 2020) ordinal regression head. Severity is ordinal: the visual progression from healthy through general to serious is ordered, and a serious-to-healthy error is more severe than a serious-to-general error. Standard cross-entropy does not encode this ordering. CORAL represents it by decomposing the three-class prediction into binary cumulative subtasks. For severity labels $s \in \{0, 1, 2\}$:

$$P(\hat{y}_s > k) = \sigma(\mathbf{w}^\top \tilde{\mathbf{f}}_{\text{sev}} + b_k), \quad k \in \{0, 1\}, \tag{3.5}$$

where $\mathbf{w} \in \mathbb{R}^D$ is a single weight vector shared across both binary subtasks and $b_k \in \mathbb{R}$ are rank-specific scalar biases. The shared weight constraint guarantees by

construction that $P(\hat{y}_s > 0) \geq P(\hat{y}_s > 1)$, which means the model is architecturally incapable of producing rank-inconsistent severity predictions. The CORAL loss is the sum of the two binary cross-entropies:

$$\mathcal{L}_{\text{CORAL}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=0}^{1}\Big[\mathbf{1}[y_s^{(i)} > k]\log\sigma(z_k^{(i)}) + \mathbf{1}[y_s^{(i)} \leq k]\log(1-\sigma(z_k^{(i)}))\Big], \quad (3.6)$$

where B is the batch size and $z_k^{(i)}$ is the logit for sample i at rank k. The shared-weight construction encourages ordinally coherent cumulative outputs. It prevents rank-inconsistent cumulative predictions, but it does not eliminate class misclassification. The CORAL-only hard prediction can be recovered as $\hat{y}_{s,\text{CORAL}} = \sum_{k=0}^{1}\mathbf{1}[\sigma(z_k) > 0.5]$; the final DCSC prediction instead uses the fused distribution defined below.

3.2.4 Disease-Conditioned Severity Prototypes

In parallel with the CORAL head, the modulated severity feature is scored against a learnable prototype matrix $\mathbf{P} \in \mathbb{R}^{K_d \times 3 \times D}$, where $\mathbf{P}[k, s, :]$ is the prototype for disease k at severity level s. The severity logit for class s is the scaled cosine similarity between the modulated feature and the disease-specific prototype:

$$z_{\text{proto}}[s] = \frac{\tilde{\mathbf{f}}_{\text{sev}} \cdot \mathbf{P}[c, s, :]}{\tau\,\|\tilde{\mathbf{f}}_{\text{sev}}\|\,\|\mathbf{P}[c, s, :]\|}, \quad s \in \{0, 1, 2\}, \quad \tau = 0.1. \quad (3.7)$$

The prototype geometry is regularised during training by an ordinal loss that enforces, for every disease k, that the healthy–general prototype distance is smaller than the healthy–serious distance, and that the general–serious distance is smaller than the healthy–serious distance:

$$\begin{aligned}\mathcal{L}_{\text{proto}} = \frac{1}{K_d}\sum_{k=1}^{K_d}\Big[&\,\text{ReLU}(\|\mathbf{P}_{k,0} - \mathbf{P}_{k,1}\| - \|\mathbf{P}_{k,0} - \mathbf{P}_{k,2}\| + m_1) \\ &+ \text{ReLU}(\|\mathbf{P}_{k,1} - \mathbf{P}_{k,2}\| - \|\mathbf{P}_{k,0} - \mathbf{P}_{k,2}\| + m_2)\Big],\end{aligned} \quad (3.8)$$

with margins $m_1 = 0.3$ and $m_2 = 0.15$, selected by grid search on the validation set. The prototype head therefore imposes an explicit ordinal geometry on the severity

feature space. Because CORAL produces two cumulative outputs while the prototype head produces three class logits, fusion is performed in a common three-class probability space. Let $q_0 = P(y_s > 0)$ and $q_1 = P(y_s > 1)$. The CORAL and prototype distributions are

$$\mathbf{p}_{\text{CORAL}} = [1 - q_0,\ q_0 - q_1,\ q_1], \qquad \mathbf{p}_{\text{proto}} = \text{softmax}(\mathbf{z}_{\text{proto}}). \tag{3.9}$$

The final distribution is

$$\mathbf{p} = \alpha_1 \mathbf{p}_{\text{CORAL}} + \alpha_2 \mathbf{p}_{\text{proto}}, \qquad \alpha_1, \alpha_2 \geq 0, \quad \alpha_1 + \alpha_2 = 1, \tag{3.10}$$

with fusion weights selected by Optuna Tree-structured Parzen Estimator (TPE) search on the validation combined score. The hard prediction is $\hat{y}_s = \arg\max_k p_k$, while the expected severity used for soft MAE is $\tilde{y}_s = \sum_{k=0}^{2} k p_k$.

3.2.5 Earth Mover's Distance Loss for Distributional Smoothing

The CORAL loss encourages rank consistency in the cumulative probabilities, but it does not explicitly encode the geometric distance between severity classes. An additional Earth Mover's Distance (EMD) loss (Hou et al., 2016) provides distributional smoothing. The EMD loss is computed on Gaussian-smoothed versions of the one-hot severity labels. For a severity label $y_s \in \{0, 1, 2\}$, the smoothed target is:

$$q_k = \frac{\exp(-(k - y_s)^2 / 2\sigma^2)}{\sum_{j=0}^{2} \exp(-(j - y_s)^2 / 2\sigma^2)}, \quad k \in \{0, 1, 2\}, \tag{3.11}$$

with $\sigma = 0.5$. The EMD loss is the squared cumulative difference between the predicted severity distribution and this smoothed target:

$$\mathcal{L}_{\text{EMD}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{k=0}^{2} \left(\sum_{j=0}^{k} (p_j^{(i)} - q_j^{(i)}) \right)^2, \tag{3.12}$$

where $p_k^{(i)}$ is the predicted probability of severity class k for sample i. EMD penalises a serious-predicted-as-healthy error more than a serious-predicted-as-general error, complementing the rank-consistency guarantee of CORAL with an explicit penalty proportional to ordinal distance.

3.2.6 Combined Loss and Training

The full DCSC training objective combines the disease classification loss, the CORAL severity loss, the prototype regularisation loss, and the EMD distributional smoothing loss:

$$\mathcal{L}_{\text{DCSC}} = \lambda_d \mathcal{L}_{\text{disease}} + \lambda_c \mathcal{L}_{\text{CORAL}} + \lambda_p \mathcal{L}_{\text{proto}} + \lambda_e \mathcal{L}_{\text{EMD}}, \tag{3.13}$$

with weights $\lambda_d = 2.0$, $\lambda_c = 1.5$, $\lambda_p = 0.3$, and $\lambda_e = 0.3$. The disease loss carries the largest weight because it provides the supervision signal for the FiLM conditioning pathway. Without an accurate disease prediction, the FiLM modulation cannot apply the correct severity-decision-boundary transformation. The CORAL weight is moderate, reflecting that ordinal supervision is important but not dominant. The prototype and EMD weights are small because they act as geometric and distributional regularisers rather than primary supervision signals. These weights were selected using an Optuna Tree-structured Parzen Estimator (TPE) search over a discretised four-dimensional weight lattice ($\lambda_d \in [1.50, 2.50]$, $\lambda_c \in [1.00, 2.50]$, $\lambda_p, \lambda_e \in [0.10, 0.60]$, in steps of 0.25 or 0.10), jointly tuned alongside learning rate, warm-up, dropout, and the FiLM curriculum schedule; 80 of the 1,260 possible weight combinations were evaluated, and the resulting configuration was held fixed across both datasets.

3.3 Experimental Results

This section presents the experimental setup, baselines, and main quantitative results on AI Challenger 2018 and the Rice Leaf Severity dataset, followed by paired statistical significance testing.

3.3.1 Datasets and Implementation

DCSC is evaluated on two complementary plant disease datasets. The AI Challenger 2018 plant disease dataset (AI Challenger, 2018) consists of 30,215 leaf images across multiple host species, disease types, and severity grades. Following the preprocessing protocol of Yang et al. (2023), the 61 original species-disease-severity combinations are rearranged into two independent label spaces: 20 disease classes (with three diseases that lack severity annotations excluded) and three ordinal severity levels. The dataset is partitioned into train, validation, and test splits in a 70/15/15 ratio using stratified sampling on joint disease-severity labels, with the split held fixed across all methods and all random seeds. This protocol is more rigorous than the 80/20 split with no held-out test set used in the original TSTC paper, and one consequence is that absolute accuracy numbers reported in this thesis differ from those in the original TSTC publication.

The Rice Leaf Severity dataset (Ritharson, 2024) contains 2,826 training images and 180 test images across four rice disease classes and three severity grades. The training set is further partitioned into train and validation subsets in an 80/20 ratio using stratified sampling on joint labels. In the multitask setting, disease accuracy reaches 100% for most modern backbones on this dataset, which means the severity grading task can be evaluated in isolation from disease classification difficulty. Figure 3.2 shows representative samples from AI Challenger 2018. The visual boundary between general and serious severity differs substantially across diseases, which is the empirical manifestation of the disease-conditional severity ambiguity DCSC is designed to address.

DCSC is implemented in PyTorch with VMamba-Small initialised from the

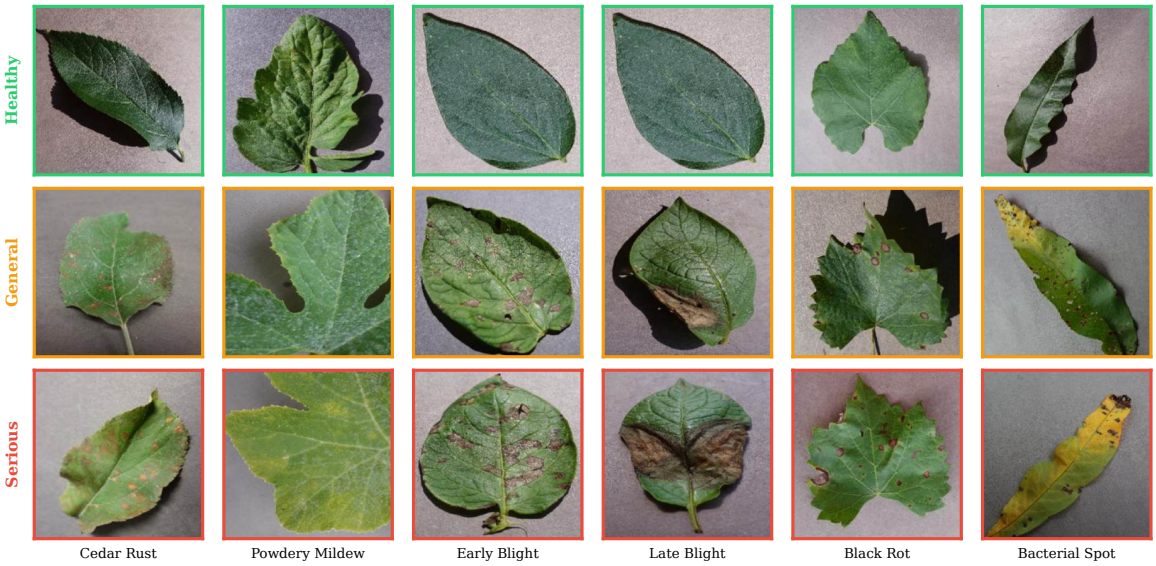


Figure 3.2: Representative samples from AI Challenger 2018 illustrating the disease-conditional nature of severity criteria. The visual boundary between general and serious severity varies substantially across disease types, which motivates the disease-conditioned modulation mechanism in DCSC.

ImageNet-1K pre-trained checkpoint and fine-tuned end-to-end. The optimiser is AdamW with a base learning rate of 2×10^{-4}, a minimum learning rate of 10^{-6}, and weight decay 0.05. The learning rate schedule is cosine annealing with 15 warmup epochs over a total of 150 training epochs. Batch size is 32, dropout is 0.1, the FiLM curriculum threshold is $T_{\text{gt}} = 20$ epochs, and the EMD smoothing parameter is $\sigma = 0.5$. Training augmentation consists of RandomResizedCrop to 224×224 pixels and RandomHorizontalFlip, followed by ImageNet mean-standard deviation normalisation. Evaluation uses a deterministic resize to 224×224 pixels with no test-time augmentation. The hyperparameter configuration was tuned on AI Challenger 2018 and applied without modification to the Rice Leaf Severity dataset, which demonstrates that the chosen setup is not overfit to a specific benchmark. All experiments are run with five random seeds (42, 123, 456, 789, 1011). Results are reported as mean and standard deviation across the five seeds. Statistical significance is tested using paired t-tests on the per-seed metric values.

3.3.2 Baselines and Evaluation Metrics

DCSC is compared against three groups of baselines. The first group contains the two strongest published multitask disease-severity models. TSTC (Yang et al., 2023) uses a Swin-Tiny backbone with three Compact Bilinear Pooling-fused branches and a five-loss deep supervision scheme. LDI-NET (Yang et al., 2024b) uses a ViT-B/16 backbone in a single-branch multi-label CNN-Transformer configuration. Both are reimplemented using the hyperparameters reported in their original papers and evaluated on the same fixed 70/15/15 split as DCSC, which means the absolute numbers reported here differ from the original publications. The second group is a VMamba baseline that holds the backbone constant and removes the DCSC innovations. This baseline uses the same VMamba-Small backbone as DCSC but replaces the FiLM-modulated CORAL severity head with two independent linear classification heads (one for disease, one for severity) trained under standard cross-entropy. The purpose is to isolate the contribution of FiLM and CORAL from any contribution arising purely from the choice of backbone. The third group is a set of two-head backbone baselines using ResNet-50 (He et al., 2016), VGG-16 (Simonyan and Zisserman, 2015), Swin-Tiny (Liu et al., 2021), and InceptionV3 (Szegedy et al., 2016), each with a shared ImageNet-pretrained backbone and two independent MLP heads. These baselines establish a backbone-diversity reference point.

Four metrics are reported throughout the experiments. Disease accuracy (DA) is the top-1 disease classification accuracy on the test set. Severity accuracy (SA) is the top-1 accuracy of the hard prediction $\hat{y}_s = \arg\max_k p_k$. The combined score is the unweighted mean of DA and SA. The reported severity mean absolute error is *soft MAE*, computed from the expected severity $\tilde{y}_s^{(i)} = \sum_{k=0}^{2} k p_{i,k}$:

$$\mathrm{MAE}_{\mathrm{soft}} = \frac{1}{N} \sum_{i=1}^{N} \left| \tilde{y}_s^{(i)} - y_s^{(i)} \right|. \tag{3.14}$$

Unlike hard-label MAE, soft MAE uses the full predictive distribution and cannot be reconstructed directly from a hard confusion matrix.

3.3.3 Results on AI Challenger 2018

Table 3.1 reports five-seed test results on AI Challenger 2018. DCSC achieves the best combined score, the best severity accuracy, and the lowest soft MAE across all evaluated methods. Disease accuracy is competitive with the strongest baseline, though as expected the differences in DA across modern backbones are small because disease accuracy is close to the ceiling on this benchmark.

Table 3.1: Five-seed test results on AI Challenger 2018 (70/15/15 fixed split). Bold indicates the best value in each column. DA = disease accuracy, SA = hard severity accuracy, Comb = unweighted mean of DA and SA, and Soft MAE is computed from the expected severity.

Method	**Backbone**	**DA (%)**	**SA (%)**	**Comb (%)**	**Soft MAE**
Proposed method					
DCSC (ours)	VMamba-S	**98.51 ± 0.23**	**92.30 ± 0.47**	**95.40 ± 0.29**	**0.091 ± 0.014**
Reproduced multitask baselines					
TSTC (Yang et al., 2023)	Swin-T	96.72 ± 0.98	90.27 ± 1.27	93.50 ± 1.09	0.099 ± 0.012
LDI-NET (Yang et al., 2024b)	ViT-B/16	97.45 ± 0.96	90.61 ± 0.82	94.03 ± 0.44	0.094 ± 0.008
VMamba baseline	VMamba-S	98.39 ± 0.22	91.01 ± 0.18	94.70 ± 0.14	0.211 ± 0.006
Two-head backbone baselines					
ResNet-50	ResNet-50	98.33 ± 0.41	90.21 ± 0.60	94.27 ± 0.38	0.099 ± 0.006
VGG-16	VGG-16	96.71 ± 0.56	90.77 ± 0.65	93.74 ± 0.49	0.094 ± 0.006
Swin-Tiny	Swin-T	97.46 ± 0.65	90.90 ± 1.21	94.18 ± 0.97	0.092 ± 0.012
InceptionV3	IncV3	97.56 ± 0.44	90.34 ± 0.86	93.95 ± 0.56	0.093 ± 0.007

The most informative comparison is the one between DCSC and the VMamba baseline. The two share the same backbone and differ only in the DCSC severity components. The VMamba baseline achieves comparable disease accuracy (98.39% versus 98.51%) and severity accuracy (91.01% versus 92.30%), but its soft severity MAE is 0.211, more than double the 0.0909 achieved by DCSC. Because soft MAE is computed from the expected severity rather than hard labels, this result indicates substantially better ordinal probability calibration; it does not imply a particular ratio of one-step to two-step hard errors. DCSC reduces soft MAE by 57% without changing the backbone.

The comparison with TSTC and LDI-NET, the two strongest published baselines, shows that DCSC improves severity accuracy by 2.03 percentage points over TSTC

(92.30% versus 90.27%) and 1.69 percentage points over LDI-NET (92.30% versus 90.61%), with MAE reductions of 8% and 3% respectively. The improvements over the published baselines are smaller in relative terms than the improvement over the VMamba baseline because TSTC and LDI-NET use ordinal-aware loss terms in their own designs, but the magnitudes still confirm that the disease-conditioned conditioning is the missing ingredient and not the backbone choice. The fixed 70/15/15 split protocol used here is more rigorous than the 80/20 split with no held-out test set used in the original TSTC and LDI-NET papers. The reproduced TSTC baseline achieves 90.27% severity accuracy on the 70/15/15 split, compared to the 88.73% originally reported on the 80/20 split. The difference reflects sample-distribution differences between the two splits and does not indicate that DCSC's improvements are an artefact of the protocol choice. The relative gap between DCSC and TSTC is preserved under both protocols.

3.3.4 Results on Rice Leaf Severity

Table 3.2 reports results on the Rice Leaf Severity dataset. Disease accuracy reaches the ceiling at 100% for DCSC, the VMamba baseline, and TSTC, which means the severity task can be analysed in isolation.

Table 3.2: Five-seed test results on the Rice Leaf Severity dataset. Bold indicates the best value in each column. Disease accuracy at ceiling for several methods isolates the severity contribution.

Method	**Backbone**	**DA (%)**	**SA (%)**	**Comb (%)**	**Soft MAE**
DCSC (ours)	VMamba-S	**100.00 ± 0.00**	**99.56 ± 0.42**	**99.78 ± 0.21**	**0.031 ± 0.007**
VMamba baseline	VMamba-S	**100.00 ± 0.00**	98.89 ± 0.35	99.44 ± 0.18	0.149 ± 0.019
TSTC (Yang et al., 2023)	Swin-T	**100.00 ± 0.00**	96.78 ± 1.08	98.39 ± 0.54	0.078 ± 0.018
LDI-NET (Yang et al., 2024b)	ViT-B/16	99.56 ± 0.42	97.33 ± 0.82	98.44 ± 0.48	0.073 ± 0.010
InceptionV3	IncV3	99.78 ± 0.27	98.67 ± 0.75	99.22 ± 0.48	0.061 ± 0.010
ResNet-50	ResNet-50	99.89 ± 0.22	98.00 ± 1.03	98.94 ± 0.59	0.068 ± 0.012
Swin-Tiny	Swin-T	99.78 ± 0.27	97.44 ± 0.75	98.61 ± 0.46	0.074 ± 0.017

DCSC achieves 99.56% severity accuracy and a soft MAE of 0.031, a reduction of 79% over the VMamba baseline (soft MAE 0.149). With disease accuracy at or near the ceiling, this comparison provides a useful isolation of the severity pathway.

The result supports a consistent benefit from disease-conditioned ordinal supervision across the two evaluated datasets.

3.3.5 Statistical Significance

Table 3.3 reports paired t-tests for the severity metrics on AI Challenger 2018. DCSC significantly outperforms both the strongest published multitask baseline (TSTC) and the VMamba baseline at $p < 0.05$ on severity accuracy, and at $p < 0.01$ on soft MAE. Disease accuracy differences are not statistically significant, which is expected given that disease accuracy is close to the ceiling for all VMamba-based methods.

Table 3.3: Paired t-test results on AI Challenger 2018 across five seeds. Asterisk indicates $p < 0.05$.

Comparison	**Metric**	t	p
DCSC vs. TSTC	Severity Acc.	3.14	0.021^*
DCSC vs. TSTC	Soft MAE	-4.51	0.003^*
DCSC vs. VMamba baseline	Severity Acc.	5.89	0.004^*
DCSC vs. VMamba baseline	Soft MAE	-13.72	$< 0.001^*$

3.4 Ablation and Analysis

This section quantifies the contribution of each DCSC component through a five-seed ablation, then examines the qualitative behaviour of the model through confusion matrix analysis, feature space visualisation, and per-disease uncertainty quantification.

3.4.1 Component Ablation

Table 3.4 presents the five-seed component ablation results on the AI Challenger 2018 dataset using the fixed 70/15/15 train–validation–test split. The final DCSC architecture combines disease-conditioned FiLM modulation, a CORAL ordinal regression head, a disease-conditioned prototype head, and Earth Mover's Distance (EMD)

regularisation. It achieves a disease accuracy of 98.51%, a severity accuracy of 92.30%, and the lowest severity MAE of 0.091.

Removing FiLM conditioning reduces disease accuracy from 98.51% to 98.02% and severity accuracy from 92.30% to 91.81%, while increasing severity MAE from 0.091 to 0.100. This degradation indicates that explicitly conditioning the severity feature space on the predicted disease identity improves the separation of disease-dependent severity levels. Without FiLM, the severity head must apply a shared ordinal decision function across all disease classes, despite differences in their visual severity criteria.

Removing the CORAL head and relying on the prototype head alone reduces severity accuracy to 91.64% and raises MAE to 0.100, while disease accuracy falls to 97.67%. The prototype head imposes ordinal geometry in the severity feature space, but without CORAL the predictions are no longer guaranteed to be rank-consistent, which is precisely the constraint that prevents the most damaging non-adjacent severity errors. This confirms that both the disease-conditioned prototype geometry and CORAL's rank-consistency constraint are necessary.

The VMamba multitask baseline obtains a comparable disease accuracy but lower severity accuracy and substantially higher severity MAE. These results indicate that the improvement of DCSC is not primarily due to disease recognition performance; instead, it arises from disease-conditioned severity modulation and ordinally informed severity optimisation.

Table 3.4: Five-seed component ablation study on AI Challenger 2018 using the fixed 70/15/15 train–validation–test split. DA and SA denote disease accuracy and severity accuracy, respectively. The best result in each column is shown in bold.

Configuration	**DA (%)**	**SA (%)**	**Soft MAE**
Full DCSC (FiLM + CORAL + Prototype + EMD)	$\mathbf{98.51 \pm 0.23}$	$\mathbf{92.30 \pm 0.47}$	$\mathbf{0.091 \pm 0.014}$
DCSC without FiLM conditioning	98.02 ± 0.28	91.81 ± 0.53	0.100 ± 0.005
DCSC without CORAL (prototype only)	97.67 ± 0.38	91.64 ± 0.58	0.100 ± 0.013
VMamba multitask baseline	98.39 ± 0.22	91.01 ± 0.18	0.211 ± 0.006

In general, the ablation results support the complementary roles of the components evaluated in this study. FiLM introduces an explicit disease-to-severity conditioning pathway, allowing the severity representation to adapt to disease-specific visual cri-

teria; CORAL enforces rank-consistent ordinal predictions; the disease-conditioned prototype head imposes ordinal geometry in the severity feature space; and EMD complements these objectives by incorporating the magnitude of ordinal prediction errors. Their combination produces the highest severity accuracy and the lowest MAE among the evaluated configurations, while retaining disease classification performance comparable to the VMamba baseline.

3.4.2 Severity Confusion Matrices

Figure 3.3 shows representative fixed-seed severity confusion matrices for the reproduced TSTC baseline and DCSC on AI Challenger 2018; Table 3.1 reports five-seed means. The TSTC matrix contains 335 serious-to-general and 187 general-to-serious errors. DCSC reduces serious-to-general errors by 52.84% (335 to 158), whereas general-to-serious errors decrease by 1.07% (187 to 185). Healthy–diseased errors are uncommon but not absent: the TSTC matrix contains 46 such errors and DCSC contains 5. These hard-count changes are descriptive and distinct from the soft-MAE significance tests.

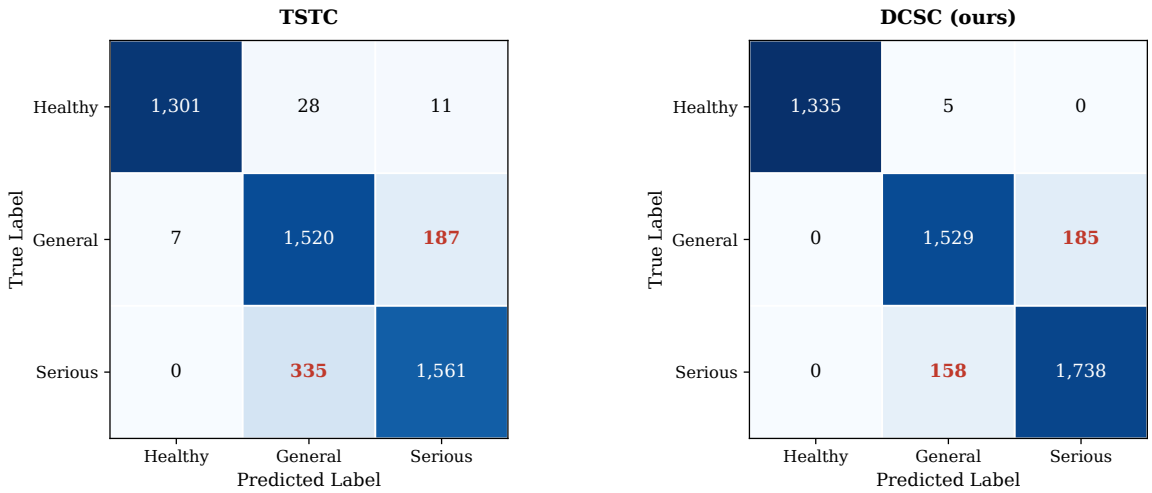


Figure 3.3: Representative fixed-seed severity confusion matrices for the reproduced TSTC baseline (left) and DCSC (right) on AI Challenger 2018. DCSC reduces serious-to-general errors from 335 to 158 (52.84%); general-to-serious errors change from 187 to 185. Table 3.1 reports five-seed means.

3.4.3 Feature Space Analysis

A central claim of DCSC is that FiLM creates disease-specific severity feature subspaces. Figure 3.4 validates this claim by projecting the severity features of all test samples to two dimensions using UMAP (McInnes et al., 2018), both before FiLM modulation (using $\mathbf{f}_{\text{sev}}$) and after FiLM modulation (using $\tilde{\mathbf{f}}_{\text{sev}}$). Two quantitative measures support the qualitative interpretation. The general-to-serious centroid distance in the projected feature space increases by 34% (from 0.570 to 0.766) after FiLM modulation. The mean within-disease general-to-serious separation increases by 40% (from 0.722 to 1.014). Both measurements indicate that the modulated feature space provides a larger geometric margin between adjacent severity classes for each disease individually, which is exactly the structure that disease-conditioned conditioning is intended to produce.

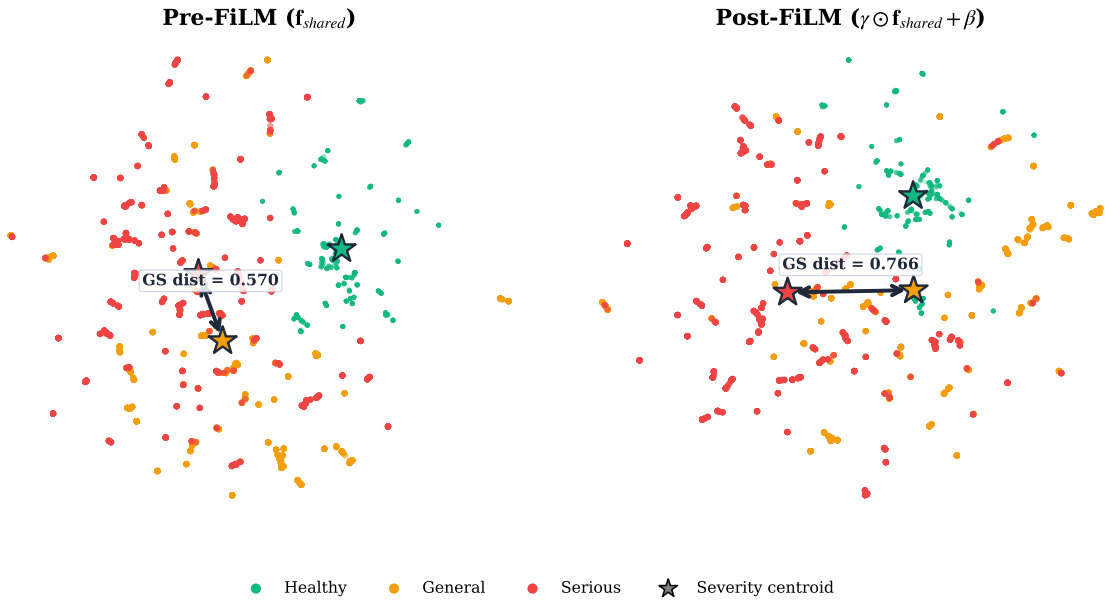


Figure 3.4: UMAP projections of severity features for all AI Challenger 2018 test samples, before and after FiLM modulation. The post-FiLM general-to-serious centroid distance increases by 34% (from 0.570 to 0.766) and the mean within-disease general-to-serious separation increases by 40% (from 0.722 to 1.014), confirming that FiLM creates disease-specific severity subspaces as designed.

3.4.4 Per-Disease Error and Uncertainty Quantification

Residual severity MAE in DCSC is not uniformly distributed across diseases. Five disease classes account for the majority of the remaining error: Puccinia Polysora (MAE 0.2146), Grape Black Rot (MAE 0.1902), Grape Black Measles (MAE 0.1796), the merged Early Blight class covering both potato and tomato hosts (MAE 0.1679), and Corn Curvularia Leaf Spot (MAE 0.1579). All five classes exhibit either continuous symptom progression where the visual boundary between general and serious is genuinely ambiguous, or multi-host annotation artefacts where the same disease label aggregates morphologically distinct presentations.

To examine whether this residual error is consistent with intrinsic label ambiguity rather than disease misclassification propagating through the FiLM pathway, MC-Dropout (Gal and Ghahramani, 2016) uncertainty quantification was applied. Thirty stochastic forward passes were performed on each of 4,950 test images with dropout active at inference time. The five high-MAE diseases showed significantly higher mean predictive entropy on the severity output than a low-MAE control set (entropy 0.1039 versus 0.0703, $p = 1.71 \times 10^{-9}$). A complementary test substituted the predicted disease identity with the ground-truth disease label during FiLM modulation. The uncertainty gap remained nearly identical under fixed ground-truth disease conditioning (entropy 0.1030 versus 0.0704, $p = 4.42 \times 10^{-9}$). This pattern is consistent with residual severity errors clustering on visually or annotationally ambiguous diseases rather than being explained solely by upstream disease misclassification. MC-Dropout entropy cannot exclude other causes, including model insufficiency or annotation noise.

3.5 Summary

This chapter introduced DCSC, a disease-conditioned severity calibration framework for plant disease multitask learning. The framework adds FiLM-based disease-conditioned modulation, CORAL ordinal supervision, disease-conditioned severity prototypes, and an Earth Mover's Distance loss to the VMamba-Small backbone. On AI Challenger 2018 under a five-seed 70/15/15 fixed-split protocol, DCSC achieves

98.51% disease accuracy and 92.30% severity accuracy with a soft MAE of 0.0909. The 57% soft-MAE reduction over the VMamba baseline is statistically significant ($p < 0.001$). In the representative confusion matrix, serious-to-general errors are 52.84% lower than for the reproduced TSTC baseline; this count reduction is descriptive and was not itself subjected to a significance test. On the Rice Leaf Severity dataset, where disease accuracy reaches the ceiling, DCSC achieves a 79% soft-MAE reduction over the VMamba baseline, supporting the benefit of the severity pathway independently of backbone choice.

The ablation study supports complementary roles for FiLM-based disease-conditioned modulation, CORAL ordinal regression, the disease-conditioned prototype head, and Earth Mover's Distance smoothing. The feature-space analysis shows increased class separation after FiLM, while the MC-Dropout results are consistent with residual errors concentrating in visually or annotationally ambiguous disease classes. The next chapter turns to the second contribution of this thesis, Knowledge-Grounded Depth-Controlled Diffusion (KGDC-Diffusion), which addresses the orthogonal problem of synthesising morphologically faithful early-stage disease images for data augmentation under class-imbalanced training conditions.

Chapter 4

Knowledge-Grounded Depth-Controlled Diffusion for Synthetic Augmentation of Plant Disease Imagery

4.1 Introduction

Deep learning architectures such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViT) have demonstrated remarkable capability in precision agriculture, yet their real-world deployment remains heavily constrained by reliance on high-quality, class-balanced datasets. In practical field conditions, agricultural datasets naturally exhibit a pronounced *long-tail* distribution. Mild-stage examples are relatively abundant, whereas Early- and Late-stage cases are less represented. Early-onset symptoms are particularly scarce because the observable window is narrow and expert botanical annotation is costly. This imbalance is visible in the SoySeverityDataset described later in this chapter: some disease–severity classes contain fewer than 40 real training images, while the largest class contains close to 500.

This scarcity cannot be resolved simply by collecting more field images: early-stage infections are transient, and a research group cannot control when or how often they occur. The alternative pursued in this chapter is generative augmentation. However, generating additional images is only useful if those images are trustworthy, that is, if their visual and statistical properties place them within the same data manifold

that the real, field-collected images occupy, rather than in some generic or unrelated region of image space. A synthetic image that looks diseased but does not resemble the feature statistics of real diseased leaves gives a downstream classifier nothing to learn from, and may actively mislead it. The central design goal of this chapter is therefore to synthesise disease–severity images whose distribution overlaps as closely as possible with the real, scarce field data, so that a classifier trained on the enlarged dataset generalises to the same kind of images it will encounter at deployment time, rather than overfitting to the narrow set of real examples available today.

To mitigate these data deficiencies, the computer vision field has pivoted towards Latent Diffusion Models (LDMs), which offer superior training stability and distributional diversity compared to traditional Generative Adversarial Networks (GANs), as discussed in Section 2.4. However, standard LDMs fail to accurately capture realistic biological variations when trained on highly imbalanced datasets. This limitation stems from their reliance on discrete class conditioning or simplistic class-name prompts; these approaches treat fluid, continuous disease progression as isolated, independent categories, causing the model to collapse distinct severity phases and struggle with subtle, evolving lesion morphologies, a failure mode identified in Section 2.4 as the *conditioning insufficiency problem*.

To resolve this without the overhead of manual spatial annotation, this chapter proposes **KGDC-Diffusion** (Knowledge-Grounded Depth-Controlled Diffusion), a two-stage generative framework optimised for plant disease imagery. Rather than conditioning generation on bare class names, the framework retrieves expert-curated morphological descriptions of each disease–severity class using a Retrieval-Augmented Generation (RAG) pipeline (Lewis et al., 2020), and injects these descriptions into the cross-attention layers of a Stable Diffusion backbone. This is paired with a depth-only ControlNet that automatically extracts geometric pseudo-depth maps via zero-shot monocular estimation, so that the generated leaf geometry stays realistic without requiring hand-drawn lesion masks or hardware depth sensors. Together, these two conditioning signals allow the framework to synthesise structurally realistic, severity-specific images that sit inside the real data manifold, are used to balance the long-tailed

SoySeverityDataset, and measurably improve downstream diagnostic classifiers.

This chapter is organised as follows. Because the retrieval-grounded prompts and the depth ControlNet both depend directly on the raw data, Section 4.2 describes the SoySeverityDataset, the botanical text corpus, and the pseudo-depth estimation procedure first, before any architectural detail is introduced. Section 4.3 then presents the proposed KGDC-Diffusion framework, covering both the RAG-guided semantic conditioning stage and the depth-controlled spatial conditioning stage. Section 4.4 describes how the two trained stages are combined at inference time into a synthesis and targeted-augmentation protocol. Section 4.5 details the experimental settings, baselines, and implementation to evaluate both the generative quality of images produced using the KGDC-Diffusion framework and its effect on image classification tasks when used for data augmentation. Section 4.6 presents the generative and downstream classification results on the SoySeverityDataset. Section 4.7 reports ablation and sensitivity studies that isolate the contribution of each pipeline component. Section 4.8 examines the computational cost and practical accessibility of the framework. Section 4.9 evaluates portability across three external datasets. Finally, Section 4.10 summarises the chapter.

4.2 Real Dataset Collection and Preprocessing

The methodology in Section 4.3 depends on two artefacts derived directly from the raw images: a botanical text corpus that supplies the descriptive language retrieved by the RAG module, and a set of pseudo-depth maps that supply the spatial condition for the ControlNet. Both are constructed from the same underlying dataset. This section is placed before the architectural description for that reason: the dataset, the corpus, and the depth maps are the foundation on which the two conditioning stages are built, and understanding what they contain makes the design choices in Section 4.3 easier to follow.

4.2.1 SoySeverityDataset

This study presents the SoySeverityDataset, a compilation of soybean leaf images containing three diseases significant in Indian agriculture: Yellow Mosaic Virus (YMV), Aerial Blight, and Bacterial Pustule. A plant pathologist at ICAR-NSRI was consulted to categorise the dataset into three severity levels, Early, Mild, and Late, based on clear morphological criteria established for each disease. The raw images were obtained from an in-house soybean disease image collection previously reported by our laboratory (Gudge et al., 2025), which provided disease-class labels but not severity annotations; the contribution of this chapter is to add expert-validated severity labels to this collection. Because annotation was validated by a single domain expert, inter-annotator agreement statistics such as Cohen's κ are not applicable, which we acknowledge as a limitation. All images are resized to 512×512 resolution to keep the computational cost of the generative pipeline manageable. The dataset additionally includes a Healthy class of 700 images, yielding a total of 2,803 annotated samples. The data is partitioned into training, validation, and test sets using stratified sampling with a 70:15:15 split; only the training partition is used for generative model training, RAG conditioning, and pseudo-depth extraction, while the validation and test partitions are reserved exclusively for downstream classifier selection and evaluation, so that no held-out image ever influences the generative pipeline. The severity-annotated images, labels, and split files are deposited on Zenodo (Vashistha, 2026) to support reproducibility. Figure 4.1 illustrates the pronounced natural class imbalance present in the field-collected data, which motivates the augmentation strategy developed in Section 4.4.

4.2.2 Botanical Text Corpus

A class-name prompt such as "mild Bacterial Pustule" tells a diffusion model almost nothing about what mild Bacterial Pustule actually looks like: it says nothing about lesion colour, shape, spread across the leaf, or texture. To give the model something more useful to condition on, this section builds a small text corpus, written

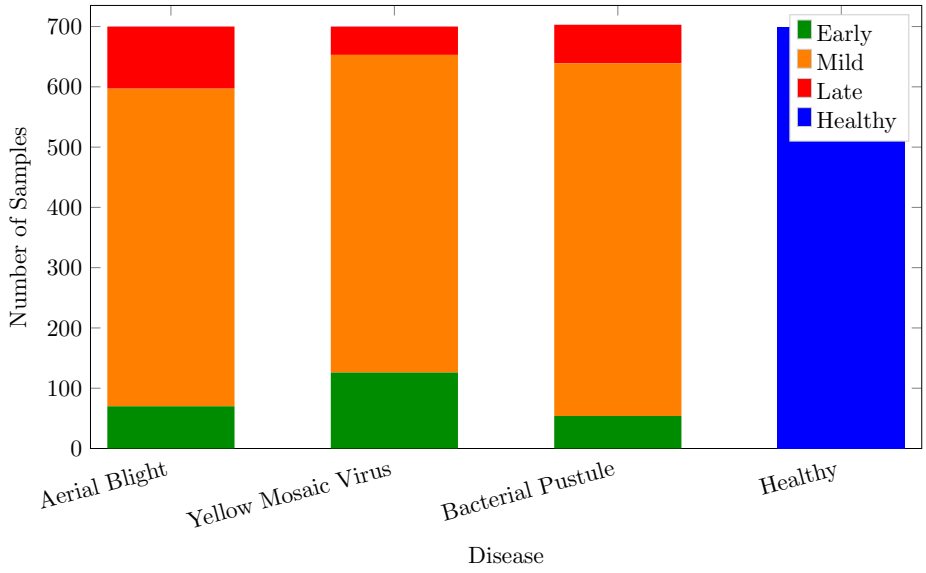


Figure 4.1: Class-wise distribution of the SoySeverityDataset across disease categories and severity levels, highlighting the natural long-tail class imbalance present in field-collected samples.

with a plant pathologist, that describes what each disease–severity class looks like in plain visual terms. This corpus is the knowledge base that Stage I retrieves from at generation time, and its construction is described here rather than in the methodology section because it does not depend on any part of the diffusion pipeline; it only depends on the disease and severity labels already established in Section 4.2.1.

For each of the ten disease–severity classes (the nine disease classes plus Healthy), we wrote eight short natural-language descriptions of 15–25 words each, kept short because CLIP ViT-L/14, the text encoder used later in Stage I, truncates any prompt beyond 77 tokens. Rather than writing eight near-identical sentences, each of the eight entries for a class was written to emphasise a different visual detail: the overall appearance, the shape of individual lesions, where the lesions sit on the leaf, the colour, the surface texture, roughly how much of the leaf is covered, what the leaf looks like when photographed in the field, and a description anchored to the underlying pathogen. Writing the entries this way, rather than as eight paraphrases of the same sentence, means that when the corpus is later embedded into a vector space (Section 4.3.1), the

eight entries for a class do not collapse to nearly the same point; retrieving different entries genuinely returns different visual information rather than a reworded copy of the same description. Table 4.1 shows two representative entries per class; the full 80-entry corpus is provided in the supplementary material.

Table 4.1: Representative entries from the botanical knowledge corpus (two per class shown). The full corpus of 80 entries is provided in the supplementary material.

Class	**Emphasis**	**Prompt Text**
Aerial Blight – Early	BASE	A soybean leaf with early symptoms of Aerial Blight, showing small water-soaked spots at the leaf tip.
Aerial Blight – Early	MARGIN	Soybean leaflet with initial Aerial Blight signs, faint grey-green patches along one margin, otherwise green.
Bacterial Pustule – Mild	BASE	A soybean leaf showing mild Bacterial Pustule, with visible pustules and slight yellowing around them.
Bacterial Pustule – Mild	TEXTURE	Mid stage Bacterial Pustule on a soybean leaf, densely speckled with raised tan pustules and faint yellow halos.
Yellow Mosaic Virus – Late	BASE	A soybean leaf with severe yellow mosaic symptoms, extensive discoloration, and deformation from late stage infection.
Yellow Mosaic Virus – Late	VEIN	Late Yellow Mosaic Virus on a soybean leaf, mostly yellow surface with thin green strips along the major veins.
Healthy	BASE	A healthy green soybean leaf with smooth texture, uniform colour, and no visible disease symptoms.
Healthy	COLOR	Disease free soybean leaflet, uniform emerald green colour, turgid and intact, ready as a healthy reference.

4.2.3 Pseudo-Depth Estimation

While the botanical corpus supplies the text side of the conditioning used in Stage I, Stage II (Section 4.3.2) additionally needs a spatial signal that tells the ControlNet

what shape a leaf takes, so that generated lesions sit on a plausible leaf surface rather than a flat or arbitrary background. Collecting this kind of information normally requires either a depth sensor or manually drawn lesion masks, neither of which is available for the SoySeverityDataset and both of which would be expensive to add retroactively. Instead, every training image is passed once through the pre-trained *DepthAnything* model (Yang et al., 2024c), which performs zero-shot monocular depth inference and produces a relative pseudo-depth map without requiring any training on our data. The resulting maps are min–max normalised to $[0, 1]$ and cached alongside the images. No manual spatial annotation, depth sensor, or lesion mask is required at any stage. These pseudo-depth maps are used directly as the spatial condition c_I that constrains the coarse structural layout of the generated leaves during Stage II inference.

4.3 Proposed Methodology

With the dataset, the botanical corpus, and the pseudo-depth maps established in Section 4.2, this section describes how KGDC-Diffusion uses them. The proposed framework comprises two functionally distinct trained stages, followed by a separate synthesis and augmentation protocol described in Section 4.4 and illustrated in Figure 4.2. Stage I addresses the semantic quality of the generated disease morphology, drawing on the botanical corpus of Section 4.2.2, and Stage II enforces geometric spatial realism, drawing on the pseudo-depth maps of Section 4.2.3.

The architecture uses Stable Diffusion v1.5 as the core generative backbone. This variant was selected for three reasons: (i) the ControlNet ecosystem for v1.5 is the most mature and well-documented, with pre-trained depth-conditioning checkpoints available; (ii) its moderate parameter count keeps training and inference within the memory budget of a single modern GPU, making the pipeline accessible to agricultural research groups without large multi-GPU clusters, a point examined more closely in Section 4.8; and (iii) using a widely adopted backbone improves reproducibility. Stage I and Stage II are each trained once and then kept frozen; how the two frozen

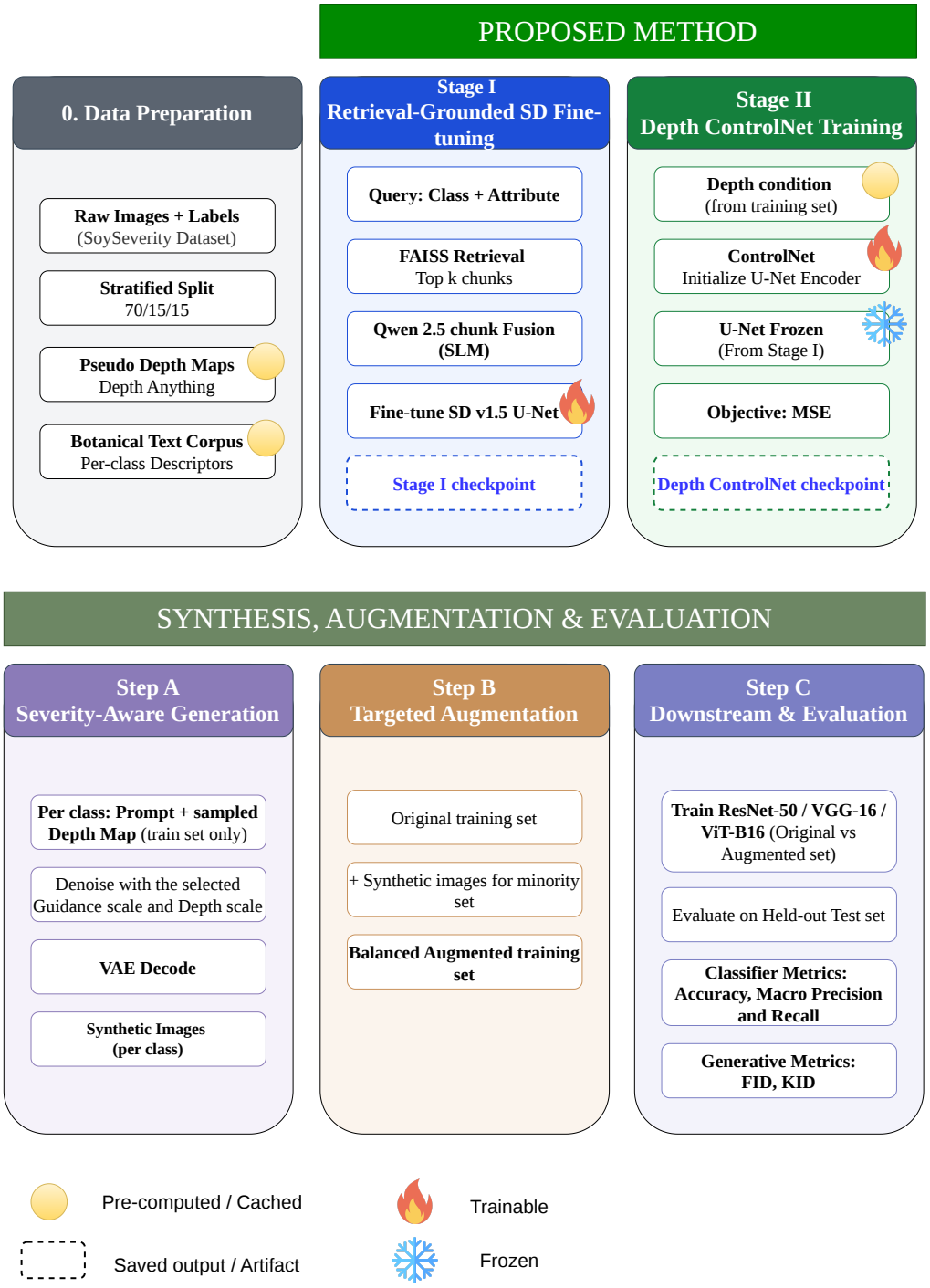


Figure 4.2: High-level block diagram depicting the overall KGDC-Diffusion framework. Stage I (RAG-guided semantic conditioning) and Stage II (depth-controlled spatial conditioning) are each trained once and then kept frozen. Steps A–C describe the synthesis, augmentation construction, and downstream evaluation protocol.

stages are subsequently combined at inference time to generate images and construct an augmented training set is described separately in Section 4.4.

4.3.1 Stage I: RAG-Guided Soft Semantic Conditioning

Bare class-name prompts give the diffusion model no information about what a disease–severity class actually looks like, and this is exactly the gap the botanical corpus of Section 4.2.2 was built to fill. Stage I closes this gap by retrieving relevant entries from the corpus and injecting them into the diffusion model's cross-attention layers, following the general Retrieval-Augmented Generation paradigm (Lewis et al., 2020), in which a generative model is grounded in text retrieved from an external knowledge source rather than relying only on its own parametric knowledge. The stage consists of two parts: retrieving and assembling a prompt from the corpus, and then using that prompt to fine-tune the diffusion model.

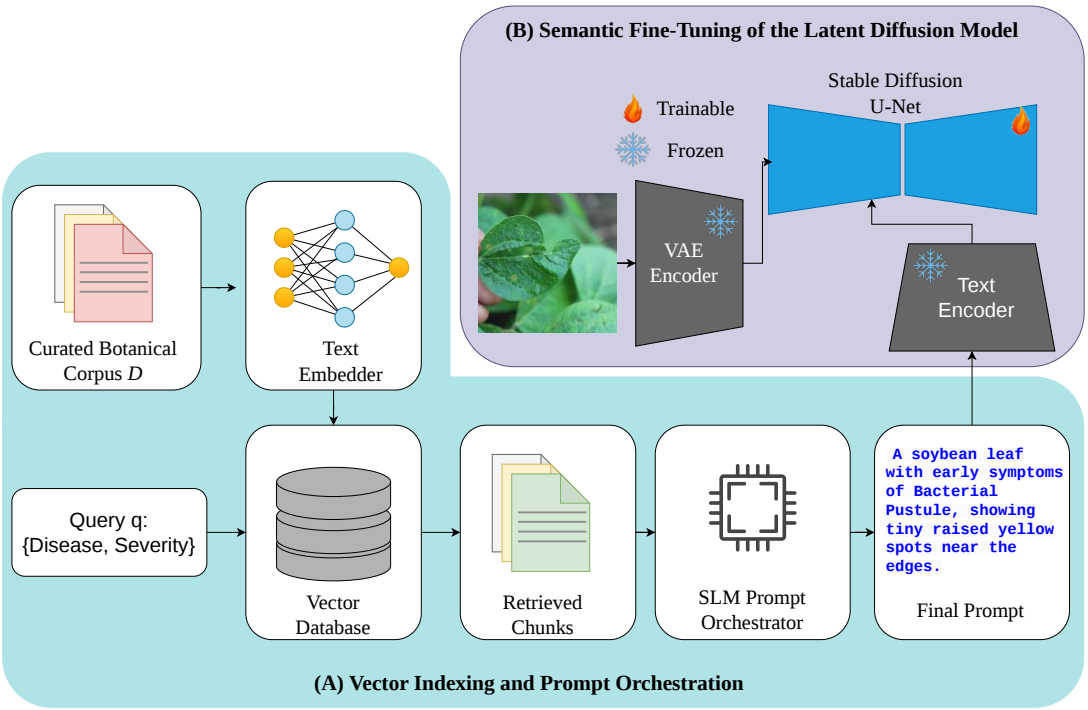


Figure 4.3: Illustration of Stage I of the proposed pipeline. Section (A) details the prompt orchestration via Retrieval-Augmented Generation (RAG); and Section (B) demonstrates the semantic fine-tuning of the Latent Diffusion Model.

4.3.1.1 Vector Indexing and Prompt Orchestration

The 80-entry botanical corpus of Section 4.2.2 is embedded into a dense vector space using the pre-trained *all-MiniLM-L6-v2* sentence encoder (Reimers and Gurevych, 2019) and indexed per class using FAISS (Johnson et al., 2019), so that retrieval for a given class is restricted to that class's own eight entries and cannot accidentally pull in descriptors from a different disease. During generation, a query is formed by combining the target class label with a randomly sampled emphasis keyword (for example, "Bacterial Pustule Mild emphasising texture"), and this query is embedded into the same vector space. Cosine-similarity retrieval then identifies the top-k most relevant entries for the target class. This retrieval procedure is illustrated in panel (A) of Figure 4.3.

The retrieved entries are not used verbatim; concatenating several short descriptions tends to produce a run-on prompt that confuses the text encoder. Instead, a small language model (SLM), *Qwen2.5-1.5B-Instruct* (Yang et al., 2024a), is prompted with few-shot examples to fuse the retrieved entries into a single, cohesive prompt of under 25 words. Keeping the prompt this short matters for two reasons: CLIP ViT-L/14 truncates anything beyond 77 tokens, and in our own generations, longer prompts tended to wash out the severity-specific detail and collapse toward a generic, intermediate-severity appearance. The resulting prompt follows a standardised template:

> *"A soybean leaf with [Severity] symptoms of [Disease], showing [Distilled Morphological Descriptors]."*

For example, a produced prompt for an early-stage infection of Aerial Blight reads: *"A soybean leaf with early symptoms of Aerial Blight, showing small brown water-soaked spots at the tip of the leaf."* Conditioning on a prompt of this form gives the text encoder concrete, expert-validated visual cues to work with, rather than a bare class name, and is the mechanism by which Stage I is intended to produce images that are both morphologically accurate and visually diverse.

4.3.1.2 Semantic Fine-Tuning of the Latent Diffusion Model

The knowledge-grounded prompt y produced above is processed by the frozen CLIP text encoder to produce embeddings $\tau_\theta(y)$, which are injected into the intermediate U-Net layers via cross-attention, as shown in panel (B) of Figure 4.3. The LDM is fine-tuned to minimise the semantic objective:

$$\mathcal{L}_{\text{Semantic}} = \mathbb{E}_{\mathcal{E}(x),\, \epsilon\sim\mathcal{N}(0,1),\, t,\, y}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|^2\right] \tag{4.1}$$

where z_t is the noisy latent representation of the ground-truth image x at time step t, and ϵ_θ is the noise predicted by the U-Net. By minimising this loss exclusively on the target dataset using the RAG-generated prompts, the LDM recalibrates its internal priors, enabling the generation of morphologically accurate and diverse early-stage samples.

Stage I governs the appearance of the disease, but it does not constrain the geometry of the leaf on which that disease appears; a lesion generated by Stage I alone can, in principle, be placed on an implausible leaf shape. Stage II, described next, addresses this remaining gap.

4.3.2 Stage II: Depth-Controlled Spatial Conditioning

Because Stage I only constrains the text-conditioned appearance of a lesion and says nothing about leaf shape or layout, Stage II introduces a second, independent conditioning signal, a depth ControlNet $\mathbf{C}_{\text{depth}}$, conditioned on the pseudo-depth maps prepared in Section 4.2.3. Using a ControlNet architecture here means the spatial signal can be injected without retraining, or risking damage to, the semantic knowledge that Stage I has already learned. The architecture of Stage II is illustrated in Figure 4.4. The ControlNet encoder is initialised from a depth-pretrained checkpoint (Zhang et al., 2023), while the semantically fine-tuned Stage I U-Net is kept entirely frozen throughout Stage II training. Spatial conditioning c_I is injected via zero-initialised convolutions $\mathcal{Z}$:

$$y_c = \mathcal{F}(x;\, \Theta) + \mathcal{Z}(\mathcal{F}(x + \mathcal{Z}(c_I);\, \Theta_{\text{copy}})) \tag{4.2}$$

where $\mathcal{F}(\cdot;\Theta)$ is the frozen Stage I U-Net backbone and Θ_{copy} represents the trainable ControlNet parameters. Because the zero-convolution connector starts at zero, the model behaves identically to the Stage I backbone at the very start of Stage II training and only gradually learns to incorporate the structural prior, so that Stage I's semantic conditioning is never disrupted. The pseudo-depth map is treated as a *relative* prior on leaf contour, surface gradient, and foreground–background organisation; no claim of metric depth or three-dimensional anatomical consistency is made.

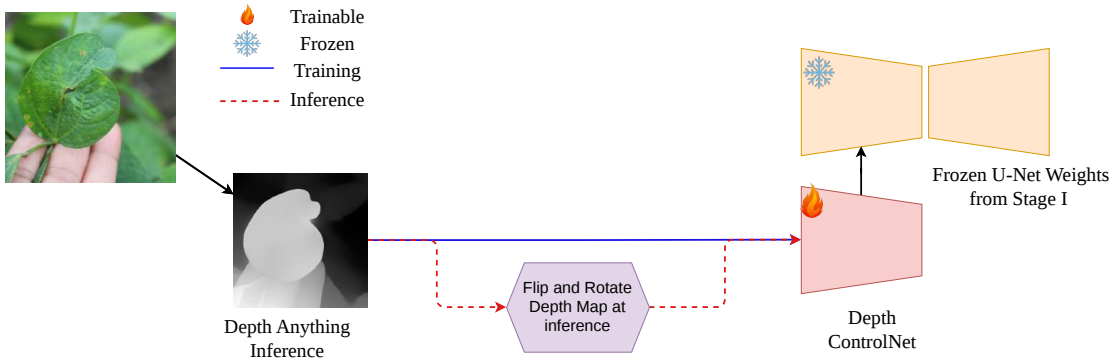


Figure 4.4: Stage II depth conditioning: a pseudo-depth map estimated from the RGB image conditions a depth ControlNet attached to the frozen, semantically fine-tuned Stable Diffusion backbone, providing coarse structural guidance over leaf geometry during generation.

At inference, for each synthetic image of a target class c, one pseudo-depth map is sampled from the training images of class c and subjected to a light spatial augmentation, a random horizontal flip or a 15°–90° rotation, before being passed to the ControlNet. As illustrated in Figure 4.5, this augmentation does more than simply reorient the leaf geometry: because the denoising process is jointly conditioned on both the depth map and the RAG-synthesised text prompt, each perturbed depth map drives the reverse diffusion along a distinct trajectory, producing outputs that vary in apparent illumination direction, colour temperature, lesion shading, and background context. The practical benefit of this is that successive samples of the same disease–severity class exhibit the kind of intra-class photometric diversity, different

lighting conditions, colour casts, and leaf surface gradients, that is essential for preventing downstream classifiers from overfitting to the narrow set of visual conditions present in the limited real training set. It also prevents the generator from reproducing memorised denoising artefacts tied to a single specific depth map, while keeping the structural prior class-consistent. How Stage I and Stage II are combined at inference time, once both are trained and frozen, is described formally as Step A in Section 4.4.

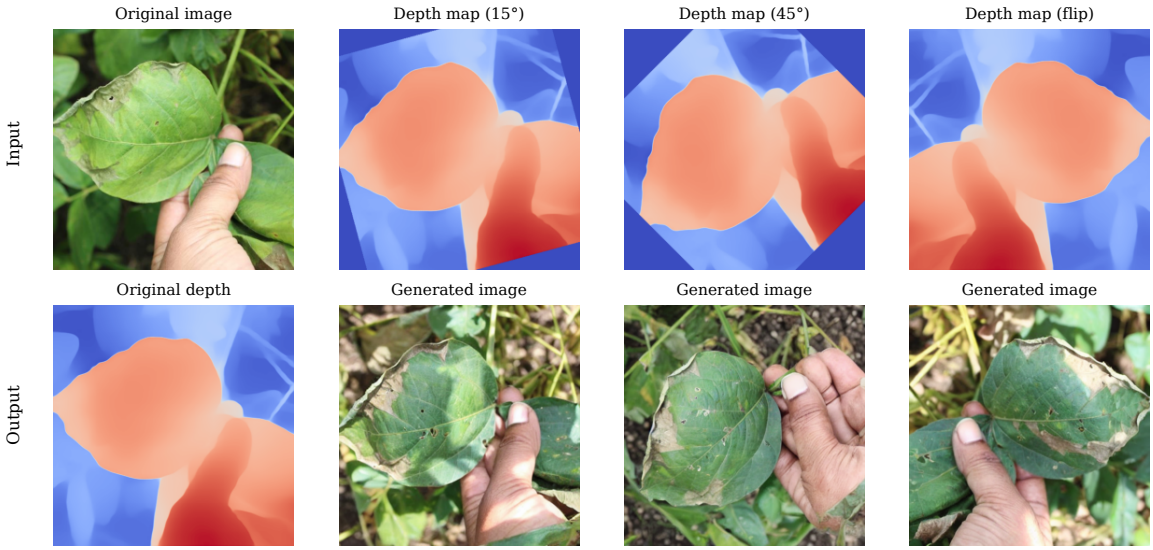


Figure 4.5: Representative pseudo-depth maps (from DepthAnything) alongside original and generated images under three augmented depth conditions (15° rotation, 45° rotation, and horizontal flip). Beyond geometric reorientation, each augmented depth map drives the denoising process along a distinct trajectory, producing outputs that differ in illumination direction, colour temperature, lesion shading, and background context, photometric diversity that is critical for downstream classifier generalisation.

4.4 Synthetic Generation and Targeted Augmentation Protocol

Stage I and Stage II, once trained, are kept entirely frozen; no further optimisation happens after this point. What remains to be specified is how the two frozen stages are operationalised, at inference time, into a supply of synthetic images, and how those images are turned into a targeted addition to the training set rather than being

added indiscriminately. This section covers both parts: Step A generates a single severity-specific image on demand, and Step B decides how many such images each class actually needs and filters out the ones that are not trustworthy enough to keep.

4.4.1 Step A: Severity-Aware Generation

Step A combines the frozen Stage I and Stage II models to synthesise a single image for a target disease–severity class c, without any further training. For every requested image, the pipeline performs three operations. First, the retrieve-then-synthesise procedure of Section 4.3.1 is re-run to produce a fresh prompt P for class c, so that successive samples of the same class draw on different emphasis aspects rather than reusing a single fixed caption. Second, a pseudo-depth map is sampled from the training images of class c and lightly augmented with a random flip or rotation, as described in Section 4.3.2, so the structural prior is class-consistent without being tied to one specific leaf instance. Third, the frozen Stage I U-Net and the Stage II depth ControlNet are jointly queried during the reverse denoising process to decode the final image.

Generation is governed by two inference-time controls that are set once for the whole pipeline rather than per image: the classifier-free guidance scale, which regulates how strongly generation follows the RAG-synthesised text prompt, and the depth-conditioning scale w_{depth}, which regulates how strongly generation follows the sampled pseudo-depth map. Based on the sensitivity analyses reported in Section 4.7, guidance scale $= 3.0$ and $w_{\text{depth}} = 0.6$ are used for all main experiments. Procedure 4.2 summarises the resulting generation process.

4.4.2 Step B: Targeted Augmentation Construction

Step A can generate images for any class on request, but generating the same number of images for every class regardless of how well-represented it already is would do nothing to correct the long-tail imbalance shown in Figure 4.1, which is the whole motivation for this chapter. Step B therefore targets generation specifically at the

Table 4.2: KGDC-Diffusion inference procedure (Step A).

Item	**Operation**
Input	Target class c, corpus $\mathcal{D}_c$, training-only depth-map pool $\mathcal{P}_c$, guidance scale s, and depth scale w_{depth}.
1	Retrieve the top-k descriptors for c from $\mathcal{D}_c$ and synthesise prompt P with the SLM.
2	Sample $c_I \sim \mathcal{P}_c$ and apply a random flip or a 15°–90° rotation.
3	Sample $\mathbf{z}_T \sim \mathcal{N}(0, I)$ and encode P and the empty prompt.
4	For $t = T, T-1, \ldots, 1$, compute conditional and unconditional noise estimates, apply classifier-free guidance, obtain the depth-ControlNet feature $\mathbf{f}_{\text{depth}}$, and update $\mathbf{z}_{t-1}$ with depth scale w_{depth}.
Output	Decode $\mathbf{z}_0$ with the VAE to obtain the synthesised image I_{syn}.

classes that need it, following minority-class augmentation and long-tailed recognition protocols (Kim et al., 2020; Park et al., 2022), and recent evidence that class imbalance itself degrades diffusion-model fidelity and diversity on tail classes (Qin et al., 2023). Let $n_{\max} = 489$ denote the size of the largest real training class (Healthy). The balancing target is set to

$$T = n_{\max} = 489, \tag{4.3}$$

and 800 candidate synthetic images are generated, using Step A, for every disease–severity class with $n_c < T$. Generating substantially more candidates than are ultimately needed is deliberate: it leaves room for a quality filter to discard the weaker samples rather than forcing every generated image into the training set regardless of quality.

This quality filter is necessary because a generative model, especially on under-represented classes, can produce images that look plausible but do not actually sit within the real data manifold, the same concern about distributional overlap raised at the start of this chapter. Following the realism-score-based cleanup strategy used for diffusion-based weed augmentation by Chen et al. (2024), itself based on the improved precision–recall formulation of Kynkäänniemi et al. (2019), each candidate image is scored against the real training manifold before being admitted. Let ϕ_g denote the feature vector of a generated image, and let Φ_r denote the set of feature vectors

extracted from all real training images. For each reference feature vector $\phi'_r \in \Phi_r$, the local neighbourhood radius is defined by the Euclidean distance to its k-th nearest neighbour within the real feature set, $\mathrm{NN}_k(\phi'_r, \Phi_r)$. The realism score of a generated image is then

$$R(\phi_g, \Phi_r) = \max_{\phi'_r \in \Phi_r} \left\{ \frac{\|\phi'_r - \mathrm{NN}_k(\phi'_r, \Phi_r)\|_2}{\|\phi_g - \phi'_r\|_2} \right\}. \tag{4.4}$$

The numerator is the local radius of the real data manifold around a given real sample ϕ'_r; the denominator is how far the generated sample ϕ_g sits from that same real sample. A generated image is retained only if $R \geq 1$ for at least one real reference point, meaning its feature representation lies inside the neighbourhood of at least one real training sample; candidates with $R < 1$ are discarded. In effect, this directly operationalises the manifold-membership justification given at the start of this chapter: only synthetic images that a real-data neighbourhood test confirms are close to genuine field data are allowed into the training set.

Table 4.3 reports the resulting augmentation plan after filtering. Consistent with long-tailed recognition protocols (Kim et al., 2020; Park et al., 2022), the largest synthetic additions occur in the most under-represented classes, but the exact number retained is set by the realism-score criterion rather than by the balancing target alone: classes for which the generator happens to produce more realistic samples retain more synthetic images, while visually more difficult classes retain fewer. This quality-controlled selection is what is used to build the augmented training set evaluated in Section 4.6.2.

Table 4.3: Per-class real-data split and realism-filtered synthetic additions for the SoySeverityDataset. The balancing target is $T = n_{\max} = 489$ (Healthy class); 800 candidate images are generated for every class with $n_c < T$ and then filtered using the realism-score criterion of Eq. 4.4.

Class	**Train**	**Val**	**Test**	**Synthetic**	**Augmented Total**
Aerial Blight – Early	49	10	11	240	289
Aerial Blight – Mild	369	79	79	120	489
Aerial Blight – Late	72	15	16	226	298
Bacterial Pustule – Early	37	8	9	430	467
Bacterial Pustule – Mild	409	88	88	80	489
Bacterial Pustule – Late	44	10	10	329	373
Yellow Mosaic Virus – Early	88	19	19	185	273
Yellow Mosaic Virus – Mild	368	79	80	121	489
Yellow Mosaic Virus – Late	32	7	8	457	489
Healthy	489	105	106	—	489
Total	**1,957**	**420**	**426**	**2,188**	**4,145**

4.5 Experimental Settings

The experimental protocol is organised to evaluate both the generative capability of KGDC-Diffusion and the practical value of the resulting synthetic data. The dataset and split configuration first establishes a consistent basis for all comparisons. The implementation details then specify the training configuration and hyperparameters required for reproducibility, followed by the generative baselines used to position the proposed method against representative alternatives. The downstream classifier protocol evaluates whether improvements in synthetic image quality translate into measurable gains in diagnostic performance. Finally, the selected metrics provide complementary measures of fidelity, diversity, distributional coverage, and downstream utility.

4.5.1 Datasets and Splits

The primary evaluation uses the SoySeverityDataset introduced in Section 4.2.1, split using the stratified 70/15/15 train/validation/test protocol described there. The generative models use only the training partition as the real reference distribution and as the source of pseudo-depth maps; validation and test images are never used for synthetic conditioning and are reserved for classifier selection and final downstream evaluation. Portability of the framework is additionally examined on three external datasets, described separately in Section 4.9; for each of these, the pipeline is retrained from scratch with a dataset-specific descriptor corpus, so those experiments test the portability of the pipeline's *design* rather than zero-shot transfer of a single trained generator.

4.5.2 Implementation Details

All experiments are conducted on a single NVIDIA H200 GPU (141 GB HBM3e) with Python 3.10 and PyTorch 2.9, using the Hugging Face Diffusers and Transformers libraries. For Stage I, the Stable Diffusion v1.5 backbone is fine-tuned at 512×512 resolution for 16,000 steps using the AdamW optimiser ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $= 1 \times 10^{-2}$), with a batch size of 8, a peak learning rate of 2×10^{-5}, a cosine learning rate scheduler with 500 warmup steps, and Min-SNR weighting ($\gamma = 5.0$) (Hang et al., 2023) for improved convergence stability. Stage II trains the depth ControlNet for 10,000 additional steps using the same optimiser, learning rate 1×10^{-5}, and a cosine schedule with 200 warmup steps, while the Stage I U-Net is kept frozen. The complete hyperparameter configuration is reported in Table 4.4.

4.5.3 Generative Baselines

KGDC-Diffusion is benchmarked against a suite of state-of-the-art generative models: StyleGAN2 (Karras et al., 2020a), StyleGAN3 (Karras et al., 2021), BigGAN (Brock et al., 2019), ADM-G with both DDIM and DDPM sampling (Dhariwal and Nichol, 2021), and a fine-tuned Stable Diffusion v1.5 baseline trained on the SoySeveri-

Table 4.4: Exhaustive list of hyperparameters used for Stage I (Base LDM) and Stage II (Depth ControlNet) training.

Hyperparameter	**Stage I (Core Diffusion)**	**Stage II (Depth ControlNet)**
Base Model	Stable Diffusion v1.5	Depth Pre-trained Checkpoint
Resolution	512×512	512×512
Optimizer	AdamW	AdamW
Adam β_1, β_2	0.9, 0.999	0.9, 0.999
Weight Decay	0.01	0.01
Learning Rate	2×10^{-5}	1×10^{-5}
LR Scheduler	Cosine	Cosine
Warmup Steps	500	200
Batch Size	8	8
Max Train Steps	16,000	10,000
Min-SNR γ	5.0	—
Noise Offset	None	None
Precision	FP16	FP16

tyDataset using simple class-name prompts, with identical hyperparameters to Stage I but without RAG conditioning. These baselines isolate three separate sources of improvement: adversarial versus diffusion-based synthesis, class-name prompting versus retrieval-grounded conditioning, and semantic-only conditioning versus semantic-plus-depth conditioning. Given the limited per-class sample sizes, particularly for the minority Early- and Late-stage classes, all GAN baselines were trained using Adaptive Discriminator Augmentation (Karras et al., 2020b) to reduce discriminator overfitting; despite this, BigGAN still exhibits severe mode collapse, reflected in its near-zero Recall in Table 4.5. Deng and Lu (2025) are not included as a direct empirical baseline because their ControlNet conditioning relies on manually annotated bounding boxes, an annotation type unavailable for the SoySeverityDataset; a like-for-like comparison is therefore not possible without fundamentally altering one of the two methodologies.

4.5.4 Downstream Classifier Training

Downstream diagnostic classifiers (VGG-16 (Simonyan and Zisserman, 2015), ResNet-50 (He et al., 2016), and ViT-B/16 (Dosovitskiy et al., 2021)) are initialised from ImageNet weights and fine-tuned for 50 epochs with batch size 32. VGG-16 and ResNet-50 use SGD with a learning rate of 1×10^{-2}; ViT-B/16 uses AdamW with a learning rate of 3×10^{-4}. To isolate the effect of the synthetic data itself, the same standard geometric transformations (random horizontal flips and minor rotations) are applied to both the baseline and augmented datasets. Each experiment is repeated across five retained random seeds to measure robustness against stochastic initialisation.

4.5.5 Evaluation Metrics

Because the framework has two goals, producing images that resemble real disease imagery and producing images that make a downstream classifier better, it is evaluated with a two-fold strategy: generative image quality and downstream classification utility which are described next.

Generative Metrics. Following standard practice (Heusel et al., 2017), generative performance is assessed along two complementary axes, fidelity (how closely a generated image resembles real images) and diversity (how much of the real variation the generated set covers). We report (i) Fréchet Inception Distance (FID) (Heusel et al., 2017), computed in the Inception-V3 feature space, as an overall fidelity measure; (ii) Kernel Inception Distance (KID) (Bińkowski et al., 2018), an unbiased alternative to FID based on squared Maximum Mean Discrepancy, useful when per-class sample counts are small; (iii) Inception Score (IS) (Salimans et al., 2016), reported for completeness only and not used for model selection, since it conflates image sharpness with inter-class diversity rather than measuring intra-class symptom coverage; (iv) Generative Precision, the fraction of synthetic samples falling within the real-data manifold; and (v) Generative Recall, the fraction of real samples covered by the generated manifold (Kynkäänniemi et al., 2019). Precision and Recall use a VGG16

feature extractor fine-tuned on the SoySeverityDataset training partition (Simonyan and Zisserman, 2015), so that the evaluation is sensitive to the subtle morphological cues, including lesion texture, chlorotic mottling, and necrotic margins, that distinguish disease severity but are not represented in generic ImageNet categories. All generative metrics use the training partition as the real reference distribution, since the validation and test partitions contain too few samples per fine-grained class for a stable estimate.

Downstream Classification Metrics. We report overall accuracy together with macro-averaged precision, recall, and F1, which weight every disease–severity class equally and are therefore sensitive to minority-class behaviour, the classes that this chapter's augmentation strategy specifically targets. Let C be the number of classes, N the total number of test samples, and TP_c, FP_c, FN_c the per-class true positive, false positive, and false negative counts:

$$\begin{aligned} \text{Accuracy} &= \frac{\sum_{c=1}^{C} TP_c}{N}, & \text{Macro-P} &= \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c}, \\ \text{Macro-R} &= \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FN_c}, & \text{Macro-F1} &= \frac{1}{C}\sum_{c=1}^{C}\frac{2P_cR_c}{P_c + R_c}. \end{aligned} \tag{4.5}$$

4.6 Results

The results are presented in two parts, following the two-fold evaluation strategy defined in Section 4.5.5. Section 4.6.1 examines whether KGDC-Diffusion generates synthetic images that are not only more visually faithful than the GAN-based, diffusion-based, and class-name-prompted baselines, but also better aligned with the distribution and feature manifold of the real data. It therefore evaluates both fidelity and coverage rather than image realism alone. Section 4.6.2 then determines whether these generative improvements translate into practical value by assessing whether targeted augmentation with the generated images improves downstream diagnostic classification, particularly for the early-stage disease classes for which real training samples are most difficult to collect.

4.6.1 Generative Image Quality

For this comparison, synthetic samples are generated only for evaluation, not passed through the realism-score filter of Section 4.4.2, and each model generates as many synthetic images per class as there are real training images in that class, so that FID, KID, IS, Precision, and Recall are computed under matched class-wise sample counts for every baseline and the proposed model.

Figure 4.6 shows a qualitative comparison of generated images across disease–severity categories. GAN-based models generate realistic leaf samples but miss severity-specific colour and lesion detail, especially for early-stage cases, which is also reflected in their tight, isolated clusters in the UMAP projections of Figure 4.7. Diffusion baselines cover the real data distribution more broadly but produce images with artefacts attributable to insufficient training samples for early-stage classes. KGDC-Diffusion performs best qualitatively, preserving subtle early-stage lesion patterns through semantically rich conditioning while generating diverse samples that extend into under-represented regions of the real data manifold.

Table 4.5 quantifies these qualitative trends. KGDC-Diffusion achieves the lowest FID (23.33) and KID (0.01293) and the highest Recall (0.951), outperforming every baseline on these three metrics. Among the non-proposed baselines, ADM-G (DDPM) is the strongest diffusion baseline (FID 35.39), while StyleGAN3 achieves the lowest FID among the GAN baselines (42.76). The fine-tuned Stable Diffusion baseline, which is identical to Stage I except that it uses class-name prompts instead of RAG conditioning, reaches only FID 57.88, which is direct evidence that fine-tuning alone is not what drives the improvement; the retrieval-grounded conditioning design does.

The Precision–Recall trade-off across architectures is instructive for the augmentation goal of this chapter. BigGAN achieves the highest Precision (0.997) but a Recall of only 0.004, indicating near-total mode collapse onto a small subset of appearances. StyleGAN2 and StyleGAN3 similarly show high Precision but cover less of the real disease manifold than the proposed model. ADM-G variants improve coverage substantially, with Recall between 0.941–0.943, at moderate Precision (0.774–0.819). KGDC-

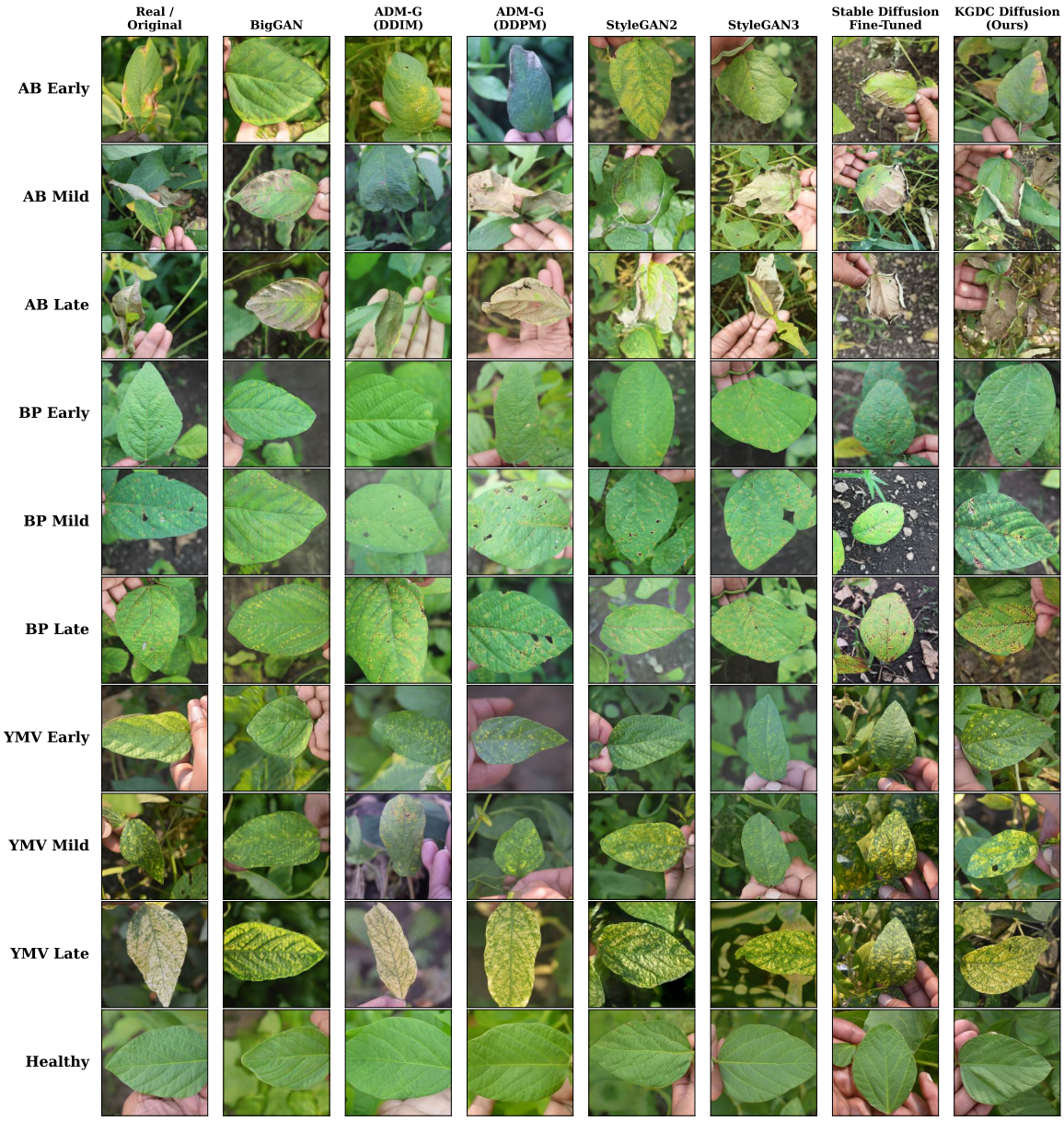


Figure 4.6: Qualitative comparison of synthetic images across representative disease–severity categories (AB: Aerial Blight, BP: Bacterial Pustule, YMV: Yellow Mosaic Virus; severity levels: Early, Mild, Late) alongside real samples. Each column corresponds to one model; each row corresponds to one disease–severity class.

Table 4.5: Quantitative comparison of images generated by different generative architectures. FID, KID, and IS are computed over the complete generated sets. Precision and Recall are computed using the dataset-trained VGG16 feature extractor. Arrows indicate the desired direction of improvement. The best result in each column is highlighted in **bold**.

Model	FID↓	KID↓	IS↑	Prec.↑	Rec.↑
BigGAN (Brock et al., 2019)	256.67	0.20020	2.11	**0.997**	0.004
StyleGAN2 (Karras et al., 2020a)	61.84	0.05230	2.87	0.929	0.826
StyleGAN3 (Karras et al., 2021)	42.76	0.01685	2.27	0.904	0.536
ADM-G (DDIM) (Dhariwal and Nichol, 2021)	53.54	0.02934	3.10	0.774	0.941
ADM-G (DDPM) (Dhariwal and Nichol, 2021)	35.39	0.01636	3.01	0.819	0.943
Stable Diffusion (fine-tuned) (Rombach et al., 2022)	57.88	0.03614	**3.25**	0.770	0.937
KGDC-Diffusion (Proposed)	**23.33**	**0.01293**	2.74	0.883	**0.951**

Diffusion achieves the highest Recall (0.951) alongside Precision of 0.883; although its Precision is lower than the three highest-Precision GAN baselines, it provides substantially broader distributional coverage while simultaneously achieving the lowest FID and KID. For minority-class data augmentation, this balance matters more than Precision alone: a classifier trained on images that cover rare and early-stage symptom configurations encounters more genuine intra-class diversity during training, which is precisely what reduces overfitting to the narrow set of visual conditions in the real data.

While Table 4.5 summarises aggregate distributional alignment, Figure 4.7 visualises the same alignment in feature space, separated into early-stage and mild/late categories so the minority early-stage classes remain visible. BigGAN forms compact, isolated clusters with limited overlap with the real distribution, consistent with the mode collapse already implied by its Recall score. KGDC-Diffusion achieves the closest alignment with the real data while also extending into sparsely populated regions of the feature space, indicating improved coverage of both common and under-represented disease variations.

4.6.1.1 Per-Class Generative Behaviour

The aggregate metrics above average over all ten classes, which can hide whether generative quality is uniform across classes or concentrated unevenly. Table 4.6 reports per-class KID rather than per-class FID, because FID is a strongly biased estimator

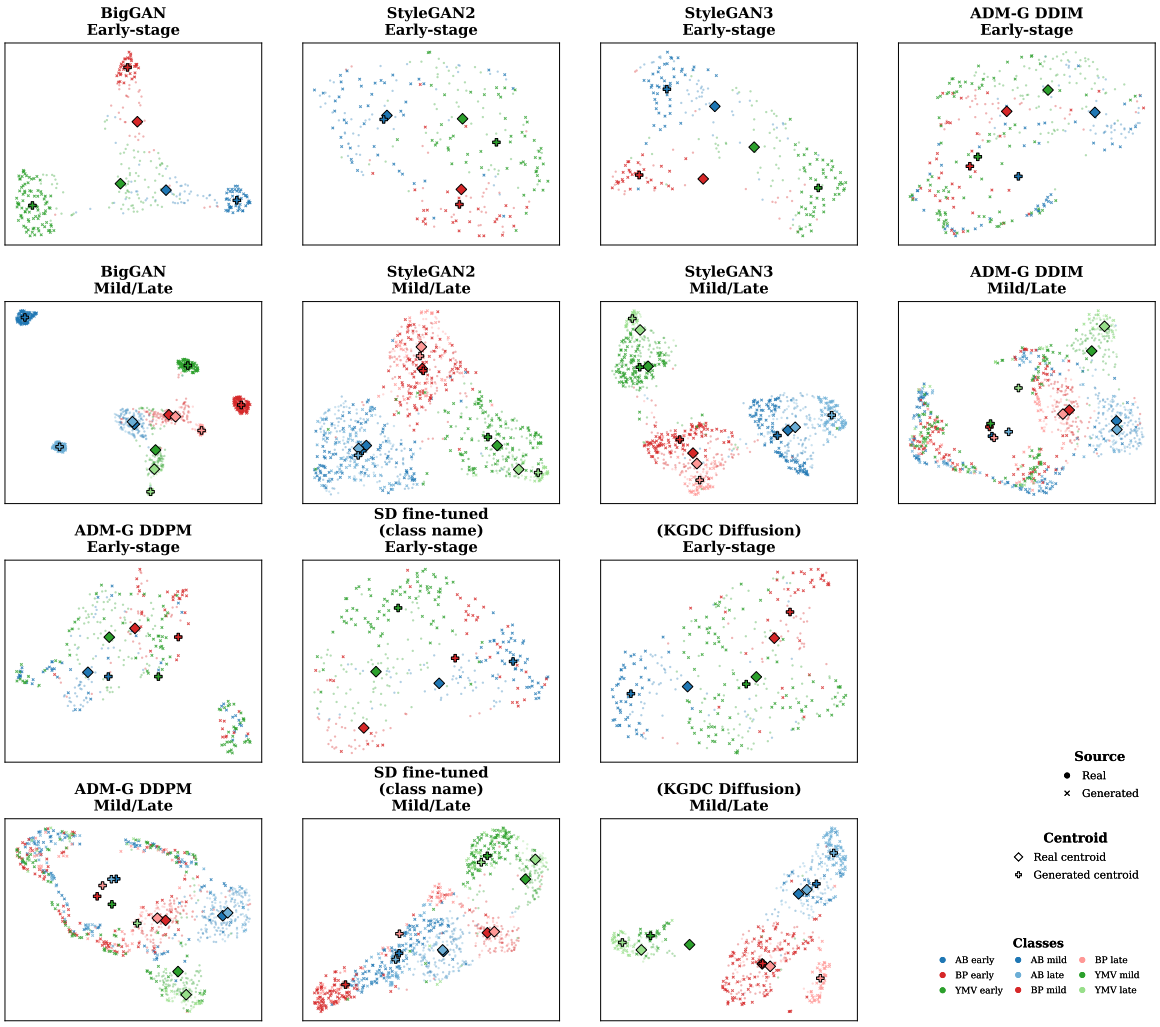


Figure 4.7: Class-stratified UMAP projections of Inception-V3 features, separating early-stage (top row) from mild/late (bottom row) disease categories across evaluated models.

at small sample sizes, and several minority classes in the SoySeverityDataset contain fewer than 100 real images.

Table 4.6: Per-class generative quality measured by KID on the SoySeverityDataset train split, using subset size $\min(50, N_c)$.

Disease & Severity Class	**Real**	**Subset**	**KID**
Healthy	489	50	0.019
Aerial Blight – Mild	369	50	0.017
Aerial Blight – Late	72	50	0.020
Aerial Blight – Early	49	49	0.023
Bacterial Pustule – Mild	409	50	0.082
Bacterial Pustule – Late	44	44	0.073
Bacterial Pustule – Early	37	37	0.072
Yellow Mosaic Virus – Mild	368	50	0.112
Yellow Mosaic Virus – Late	32	32	0.070
Yellow Mosaic Virus – Early	88	50	0.056

KID is lowest for the Healthy and Aerial Blight classes. Yellow Mosaic Virus–Mild records the highest per-class KID (0.112) despite having 368 real training samples, one of the largest classes in the dataset, which is a useful counter-example to the assumption that generative difficulty tracks sample scarcity: here the within-class visual heterogeneity of Yellow Mosaic Virus, a wide range of chlorotic mottling patterns and leaf deformation severities at the mild stage, is the dominant difficulty rather than the number of training images available. In short, the per-class breakdown shows that the framework's generative quality is not uniform, and that where it is weakest is governed more by how visually diffuse a class's symptoms are than by how many real examples exist for it, a distinction worth keeping in mind when interpreting the downstream classification results in Section 4.6.2.

4.6.2 Downstream Classification on SoySeverityDataset

Having established that KGDC-Diffusion produces higher-fidelity, higher-coverage synthetic images than the compared baselines, this section evaluates whether that

generative advantage translates into better diagnostic classification, particularly for the Early-stage classes that are both the rarest and the most agronomically valuable to identify early. Synthetic images were generated for under-represented classes, filtered using the realism-score criterion of Eq. 4.4, and added according to the quality-controlled plan in Table 4.3. This augmentation is intended to complement, not replace, continued field data collection.

Table 4.7 reports macro-averaged accuracy, precision, recall, and F1 for VGG-16, ResNet-50, and ViT-B/16, trained on the original versus the realism-filtered augmented dataset, over five retained random seeds. All three architectures show consistent improvements across every metric, together with a substantial reduction in seed-to-seed variance. ResNet-50 achieves the highest overall accuracy after augmentation (91.41%), reducing its cross-seed standard deviation from 2.55 to 0.81. VGG-16 shows the largest gains in accuracy, macro-precision, and macro-F1 (+3.76, +7.11, and +5.59 percentage points, respectively), while ViT-B/16 shows the largest macro-recall gain (+5.87 percentage points). The consistent variance reduction across all three architectures is itself a practically important result for agricultural field deployments, where compute constraints often restrict training to a single run and prevent post-hoc selection of the best seed.

4.6.2.1 Statistical Significance and Seed Stability

To confirm that these gains reflect a genuine effect of augmentation rather than favourable seed variation, paired t-tests are run over the five retained seeds, comparing per-seed accuracy with and without augmentation for each architecture. Table 4.8 reports the results: the accuracy improvement is statistically significant at the 0.05 level for all three architectures, with p-values ranging from 0.0021 to 0.0325 and 95% confidence intervals that exclude zero in every case.

Table 4.7: Downstream classification performance (mean ± standard deviation over five retained seeds) for each architecture trained without and with realism-filtered targeted synthetic augmentation. Precision, Recall, and F1 are macro-averaged over the ten disease–severity classes. Bold in the Δ rows marks the largest gain in each metric across architectures.

Condition	**Model**	**Accuracy (%)**	**Macro-P (%)**	**Macro-R (%)**	**Macro-F1 (%)**
Original	ResNet-50	88.08 ± 2.55	80.52 ± 4.65	79.78 ± 2.08	79.51 ± 3.19
	VGG-16	87.23 ± 1.16	77.16 ± 2.60	76.95 ± 1.07	76.29 ± 1.86
	ViT-B/16	87.09 ± 1.39	78.30 ± 3.14	75.98 ± 1.78	76.43 ± 1.92
With Augmentation	ResNet-50	91.41 ± 0.81	85.16 ± 1.42	82.91 ± 1.30	83.75 ± 1.27
	VGG-16	90.99 ± 0.49	84.27 ± 1.33	80.87 ± 1.22	81.89 ± 1.20
	ViT-B/16	89.44 ± 0.72	81.73 ± 1.47	81.85 ± 0.75	81.30 ± 0.66
Δ (Aug. − Orig.)	ResNet-50	+3.33	+4.65	+3.13	+4.24
	VGG-16	**+3.76**	**+7.11**	+3.92	**+5.59**
	ViT-B/16	+2.35	+3.44	**+5.87**	+4.87

Table 4.8: Paired t-test results for classification accuracy (original vs. augmented), $n = 5$ retained seeds per model.

Model	**Mean Δ Acc. (%)**	**95% CI**	p**-value**
ResNet-50	+3.33	[+0.45, +6.21]	0.0325
VGG-16	+3.76	[+2.28, +5.23]	0.0021
ViT-B/16	+2.35	[+1.43, +3.27]	0.0021

4.6.2.2 Minority-Class and Early-Stage Effects

Because the augmentation strategy specifically targets under-represented classes, the macro-averaged and class-wise breakdown is more informative than the aggregate accuracy figures above for judging whether the augmentation did what it was designed to do. Table 4.9 breaks down the performance of the best-performing model, ResNet-50, by individual disease–severity class. The largest improvements are concentrated in the classes that Table 4.3 allocated the most synthetic images to: Aerial Blight–Early (F1: 71.50 → 84.08), Aerial Blight–Late (F1: 57.08 → 61.87), and Yellow Mosaic Virus–Early (F1: 67.07 → 73.37). This is direct evidence that the synthesised images

introduce realistic intra-class variation precisely where it is scarce in the real, field-collected data, which is the outcome the targeted augmentation rule in Section 4.4.2 was designed to produce. A small number of classes show a modest recall decrease after augmentation, but these are accompanied by precision or F1 gains in most cases, suggesting a shift toward more conservative but more reliable class boundaries rather than a uniform push for higher recall everywhere.

Table 4.9: Class-wise classification performance of ResNet-50 before and after realism-filtered targeted synthetic data augmentation, averaged over five retained seeds.

	Before Augmentation (Baseline)			**After Augmentation (KGDC-Diffusion)**		
Class	**Prec. (%)**	**Rec. (%)**	**F1 (%)**	**Prec. (%)**	**Rec. (%)**	**F1 (%)**
Aerial Blight – Early	79.06	65.45	71.50	**83.06**	**85.45**	**84.08**
Aerial Blight – Mild	85.45	87.59	86.47	**88.87**	**92.66**	**90.70**
Aerial Blight – Late	56.63	**58.75**	57.08	**69.63**	56.25	**61.87**
Bacterial Pustule – Early	69.58	**73.33**	69.99	**75.90**	71.11	**72.89**
Bacterial Pustule – Mild	93.15	89.77	91.39	**93.88**	**93.86**	**93.86**
Bacterial Pustule – Late	68.91	74.00	71.12	**75.89**	74.00	**74.62**
Yellow Mosaic Virus – Early	60.46	**77.89**	67.07	**71.41**	75.79	**73.37**
Yellow Mosaic Virus – Mild	92.11	86.00	88.76	**93.00**	**92.50**	**92.73**
Yellow Mosaic Virus – Late	100.00	85.00	91.81	100.00	**87.50**	**93.33**
Healthy	99.81	100.00	99.91	**100.00**	100.00	**100.00**

Taken together, Sections 4.6.1 and 4.6.2 show a consistent story: KGDC-Diffusion produces synthetic images that are both closer to the real data manifold, by FID, KID, and Recall, and demonstrably useful once fed into downstream training, with statistically significant gains concentrated in exactly the under-represented, early-stage classes the framework targets. The next section isolates which parts of the framework are responsible for these gains.

4.7 Ablation and Sensitivity Analysis

The results above evaluate KGDC-Diffusion as a whole. This section takes it apart, to check whether the gains come from fine-tuning in general, from the retrieval-grounded prompting specifically, from the number of descriptors retrieved, or from the depth conditioning, and to check whether the inference-time settings selected in Section 4.4.1 (guidance scale 3.0, $w_{\text{depth}} = 0.6$) are a robust choice or a narrow, overtuned one.

4.7.1 Prompt Conditioning Ablation

To isolate the contribution of the retrieval mechanism itself, class-name prompting is compared against retrieval-grounded prompting using $k = 1$ through $k = 6$ retrieved descriptors, with every other pipeline component held fixed. As shown in Table 4.10, retrieval with $k = 2$ gives the best trade-off among the retrieval-grounded settings, with the lowest FID (42.27) and KID (0.0273) and the highest Precision. The class-name baseline has higher Recall, but at substantially worse FID and KID, echoing the fidelity–coverage trade-off already seen in Table 4.5. Increasing k beyond two introduces less relevant descriptors that dilute the conditioning signal and degrade FID, confirming that more retrieved text is not automatically better; the purpose of retrieval is to add specific, relevant detail rather than volume. These values are reported at a common reference guidance scale; the selected $k = 2$ configuration is then separately tuned over the guidance scale in Table 4.12.

Table 4.10: Prompt-conditioning ablation on the SoySeverityDataset at a common reference guidance scale. Class-name is the fine-tuned SD baseline; retrieval is evaluated for $k = 1$ to 6.

Prompt Strategy	**FID ↓**	**KID ↓**	**IS ↑**	**Precision ↑**	**Recall ↑**
Class-name (fine-tuned SD)	57.88	0.0361	**3.25**	0.770	**0.937**
Retrieval k=1	50.18	0.0325	2.86	0.7904	0.9043
Retrieval k=2 (Proposed)	**42.27**	**0.0273**	3.01	**0.8229**	0.9193
Retrieval k=3	48.31	0.0304	3.14	0.8164	0.9149
Retrieval k=4	49.13	0.0307	3.05	0.8104	0.8992
Retrieval k=5	54.43	0.0351	3.03	0.7987	0.9036
Retrieval k=6	54.15	0.0340	2.99	0.7909	0.9060

4.7.2 Component Ablation

To quantify the contribution of each stage in the proposed pipeline, components are introduced incrementally, beginning with the base Stable Diffusion model. As summarized in Table 4.11, the unadapted model performs poorly on this specialized

agricultural domain, yielding an FID of 224.67 and a Precision of 0.61. This result indicates that representations learned from general natural-image data do not transfer adequately to fine-grained disease-severity generation without domain-specific adaptation.

Fine-tuning the model on the target dataset using class-name prompts substantially improves performance, reducing the FID to 57.88. Replacing these generic class labels with retrieval-grounded prompts using $k = 2$ and a guidance scale of 3.0 further lowers the FID to 38.89, demonstrating that expert botanical descriptions provide more informative semantic conditioning than class names alone. Incorporating the Stage II depth ControlNet produces the strongest overall result, reducing the FID to 23.33 and increasing Recall to 0.951. These findings show that structural guidance derived from leaf geometry complements the semantic gains obtained through retrieval-grounded conditioning rather than duplicating them.

Table 4.11: Component ablation of the KGDC-Diffusion framework. Each row adds one component to the previous configuration. Metrics are computed against the training-set reference (Inception-V3 for FID/KID; dataset-trained VGG16 for Precision/Recall).

Configuration	**RAG**	**Depth**	**FID ↓**	**KID ↓**	**Precision ↑**	**Recall ↑**
Stable Diffusion (base)			224.67	0.174	0.61	0.35
SD fine-tuned (class-name prompts)			57.88	0.036	0.770	0.937
Retrieval-grounded prompting	✓		38.89	0.024	0.800	0.910
KGDC-Diffusion (RAG + Depth)	✓	✓	**23.33**	**0.01293**	**0.883**	**0.951**

4.7.3 Guidance-Scale Sensitivity

Having established that both components contribute independently, the remaining question is whether the specific operating point chosen for them, guidance scale 3.0 and $w_{\text{depth}} = 0.6$, is a stable choice or a narrowly tuned one. Table 4.12 sweeps the classifier-free guidance scale, with the depth ControlNet not yet applied, and shows a clear minimum FID at a guidance scale of 3.0. Higher guidance scales increase Inception Score but steadily reduce Recall, which is consistent with the general effect

of guidance: pushing the model to follow the text prompt more strongly narrows the generated distribution and reduces coverage of the real disease variation, so guidance scale 3.0 is used as a genuine trade-off point rather than an extreme setting.

Table 4.12: Guidance-scale sensitivity for the retrieval-grounded model (without depth ControlNet) on the SoySeverityDataset, evaluated against the training-set reference. Precision and Recall are computed in the dataset-trained VGG16 feature space.

Guidance Scale	**FID ↓**	**KID ↓**	**IS ↑**	**Precision ↑**	**Recall ↑**
1.0	61.14	0.052	2.84	0.77	0.91
2.0	43.24	0.030	2.83	0.79	0.91
3.0	**38.89**	**0.024**	2.88	0.80	0.91
4.0	39.75	0.024	2.97	0.82	0.90
5.0	42.50	0.026	3.02	0.82	0.88
6.0	45.69	0.029	3.02	0.82	0.86
7.0	49.11	0.032	3.10	0.82	0.86
8.0	52.00	0.035	3.12	0.82	0.86
9.0	54.72	0.037	3.18	0.82	0.85
10.0	56.94	0.039	3.13	0.82	0.83

4.7.4 Depth-Conditioning Scale Sensitivity

To examine the influence of the depth ControlNet conditioning weight, we vary w_{depth} while keeping the guidance scale fixed at 3.0. As reported in Table 4.13, $w_{\text{depth}} = 0.6$ provides the best overall performance, achieving the lowest FID of 23.33 and KID of 0.01293, together with the highest Precision and Recall. This setting is therefore adopted for the full KGDC-Diffusion system throughout this chapter. Increasing the depth-conditioning weight beyond 0.6 progressively over-constrains the diffusion process, leading to a slight deterioration in FID and a reduction in Recall. These results suggest that pseudo-depth is most effective when used as a coarse structural prior rather than as a rigid geometric constraint.

Together, the four ablations in this section show that neither the retrieval-grounded prompting nor the depth conditioning is redundant, and that the chosen inference-time settings sit at a genuine, non-marginal optimum rather than at the edge of a narrow

Table 4.13: Depth ControlNet conditioning-scale sensitivity (guidance scale fixed at 3.0). Precision and Recall are computed in the dataset-trained VGG16 feature space; ± values indicate variation across repeated synthetic sampling/evaluation runs.

w_{depth}	**FID ↓**	**KID ↓**	**Precision ↑**	**Recall ↑**
0.4	29.74	0.017	0.85 ± 0.11	0.94 ± 0.10
0.6	**23.33**	**0.012**	$\mathbf{0.88 \pm 0.10}$	$\mathbf{0.95 \pm 0.08}$
0.8	23.38	0.012	0.88 ± 0.10	0.95 ± 0.09
1.0	23.86	0.013	0.87 ± 0.10	0.94 ± 0.09
1.2	24.68	0.014	0.87 ± 0.11	0.93 ± 0.09

tuning window. Section 4.8 now turns from accuracy to a different, equally practical question: what this framework costs to run.

4.8 Computational Cost and Practical Feasibility

A generative pipeline is only useful to agricultural research groups if it can actually be run on the hardware such groups typically have access to; this section reports what KGDC-Diffusion costs to train and run, and checks whether that cost is confined to high-end accelerators. Table 4.14 reports the per-image inference cost and Table 4.15 the per-step training cost for each generative model on a single H200 GPU. For inference, latency and FLOPs are reported per generated 512×512 image, with FLOPs for diffusion models accumulated over all sampling steps; for training, the per-step latency, throughput, and peak memory are reported at batch size 8.

As expected, diffusion-based generators, including ours, incur substantially higher per-image inference cost than single-pass GANs, since they require many iterative denoising steps rather than one forward pass. The depth ControlNet adds a modest overhead over the Stage I backbone alone, growing from 860M to 1.22B parameters at inference, while the retrieval and depth-estimation steps are one-time preprocessing costs amortised across subsequently generated images. The runtime figures in Tables 4.14 and 4.15 were measured on a single NVIDIA H200 to keep comparisons consistent across baselines. Separately, both Stage I and Stage II fine-tuning were ver-

Table 4.14: Inference compute cost per model at 512×512, batch size 8. FLOPs for diffusion models are total over all sampling steps.

Model	**Variant**	**Params**	**Steps**	**FLOPs/image**	**Latency/image (ms)**
StyleGAN	StyleGAN2	30.5M	1	84.0 GF	2.36
	StyleGAN3	25.1M	1	545 GF	5.06
BigGAN	BigGAN	9.8M	1	27.4 GF	1.54
ADM	DDIM	613M	50	209 TF	2,093
	DDPM	613M	1000	4,172 TF	41,705
SD1.5	Class-name	860M	40	27.1 TF	314
	RAG k=2	860M	40	27.1 TF	315
	Depth ControlNet (Ours)	1.22B	40	36.4 TF	458

Table 4.15: Training compute cost per optimiser step at 512×512, batch size 8. ADM rows reflect a single generative-model step. The SD1.5 ControlNet row is ControlNet-only (U-Net frozen).

Model	**Variant**	**Params (train)**	**Latency/image (ms)**	**Peak VRAM (GiB)**
StyleGAN	StyleGAN2	61.9M	19.8	9.06
	StyleGAN3	56.5M	32.4	12.10
BigGAN	BigGAN	20.7M	21.0	7.01
ADM	DDIM	559M	146.8	112.9
	DDPM	559M	130.9	112.9
SD1.5	Class-name	860M	81.2	16.78
	RAG k=2	860M	76.9	16.78
	Depth ControlNet (Ours)	361M	106.7	10.67

ified with the same pipeline on a single NVIDIA L4 accelerator with 24 GB VRAM. The L4 is not a consumer GPU; this verification tests whether the pipeline fits within a 24-GB memory budget comparable to that available on high-end consumer hardware. The result supports single-accelerator feasibility without implying that the H200 or L4 is consumer hardware.

4.9 Cross-Dataset Portability

Every result so far comes from a single dataset, SoySeverityDataset, so it remains an open question whether the conditioning design generalises or whether it happens to fit the particular crop, disease taxonomy, and field-imaging conditions of soybean. This section checks portability on three external datasets that differ from SoySeverityDataset in crop, taxonomy, and acquisition environment: a PlantVillage tomato subset (Hughes and Salathé, 2015) under controlled imaging conditions, and two real-field weed-identification datasets, CottonWeedID10 and CottonWeedID15 (Chen et al., 2022), together with DeepWeeds (Olsen et al., 2019). In every case, the pipeline, including the descriptor corpus and the feature extractor used for evaluation, is retrained from scratch on the target dataset; these experiments therefore test the portability of the pipeline *design*, not zero-shot or few-shot transfer of a single trained generator.

4.9.1 Controlled-Environment Portability: PlantVillage Tomato

On the PlantVillage tomato subset (Table 4.16), which is not a severity-labelled dataset and therefore tests only class-level portability, the depth-controlled model reduces FID from 30.20 to 21.77 and raises Precision from 0.7609 to 0.8123 over the class-name baseline. Because PlantVillage's backgrounds are controlled and near-uniform, this is a less challenging imaging regime than field-collected soybean data, so the result should be read as evidence of pipeline portability in a controlled setting rather than as evidence of severity-aware generalisation.

Table 4.16: External validation on the PlantVillage tomato subset. The depth-controlled model substantially improves over the fine-tuned class-name baseline.

Model	FID ↓	IS ↑	Precision ↑	Recall ↑
BigGAN	54.35	2.37	0.7864	0.2202
StyleGAN2	27.63	2.67	0.7233	0.4564
StyleGAN3	30.05	2.27	0.7044	0.4902
ADM-G (DDIM)	57.71	3.95	0.5115	0.4053
ADM-G (DDPM)	34.45	2.88	0.6415	0.6885
Stable Diffusion (fine-tuned)	30.20	3.43	0.7609	0.6412
KGDC-Diffusion (Proposed)	**21.77**	3.01	**0.8123**	**0.7064**

4.9.2 Real-Field Generative Portability: CottonWeedID10 and DeepWeeds

Generative performance is evaluated on two real-field weed-species datasets: CottonWeedID10 and DeepWeeds. As shown in Tables 4.17 and 4.18, KGDC-Diffusion achieves the best FID (21.76), Inception Score (9.36), and Precision (0.6858) on CottonWeedID10, whereas ADM-G (DDPM) obtains the highest Recall (0.8035). A similar trend is observed on DeepWeeds, where KGDC-Diffusion records the lowest FID (30.67), highest Inception Score (4.28), and highest Precision (0.8743), while ADM-G (DDPM) retains the best Recall (0.6395).

Table 4.17: Generative performance on the CottonWeedID10 weed-species dataset. The best value in each column is highlighted in **bold**.

Model	FID ↓	IS ↑	Precision ↑	Recall ↑
BigGAN	107.22	5.38	0.5605	0.4315
StyleGAN2	43.90	7.39	0.6850	0.4430
StyleGAN3	26.28	5.81	0.5555	**0.7975**
ADM-G (DDIM)	40.76	5.05	0.6130	0.7300
ADM-G (DDPM)	24.14	6.70	0.6805	0.8035
Stable Diffusion (fine-tuned)	37.59	7.11	0.4847	0.5809
KGDC-Diffusion (Proposed)	**21.76**	**9.36**	**0.6858**	0.6875

Table 4.18: Generative performance on the DeepWeeds weed-species dataset. The best value in each column is highlighted in **bold**.

Model	FID ↓	IS ↑	Precision ↑	Recall ↑
BigGAN	197.18	1.77	0.3940	0.0285
StyleGAN2	44.37	1.81	0.5975	0.5565
ADM-G (DDIM)	56.02	2.41	0.5185	0.5985
ADM-G (DDPM)	38.17	3.25	0.7065	**0.6395**
Stable Diffusion (fine-tuned)	39.24	4.31	0.6762	0.4358
KGDC-Diffusion (Proposed)	**30.67**	**4.28**	**0.8743**	0.6297

Across all three external datasets, the proposed conditioning consistently improves fidelity, FID and Precision, over the compared baselines, mirroring the pattern already established on soybean. This confirms that the gains stem from the conditioning design itself rather than from properties specific to soybean imagery: the same retrieve-then-generate, depth-guided recipe transfers cleanly to a controlled-background fruiting crop and to two independent real-field weed datasets, even though each requires its own separately trained model and descriptor corpus.

4.9.3 Cross-Dataset Downstream Validation: CottonWeedID15

The three portability checks above establish that the conditioning design improves generative fidelity on external data, but they do not, by themselves, show that this generative advantage translates into a better downstream classifier outside the soybean severity task, the same gap that motivated Section 4.6.2 for SoySeverityDataset. This subsection closes that gap using the full 15-class release of CottonWeedID (Chen et al., 2022), which, unlike the 10-class subset used for the generative-fidelity comparison above, includes several minority classes with substantially fewer real images, giving a real-field, class-imbalanced setting distinct in crop, taxonomy, and imaging conditions from soybean.

Synthetic images were generated to selectively oversample the under-represented

classes of the CottonWeedID15 training partition, following the same augmentation protocol used for SoySeverityDataset in Section 4.4.2, and filtered with the same realism-score criterion. ResNet-50, VGG16-BN, and ViT-B/16 were then trained on both the real-only and the real-plus-filtered-synthetic datasets. Table 4.19 reports the results.

Table 4.19: Downstream classification performance on CottonWeedID15 (mean ± standard deviation over five seeds) for each architecture trained without and with filtered targeted synthetic augmentation.

Condition	**Model**	**Accuracy (%)**	**Macro-P (%)**	**Macro-R (%)**	**Macro-F1 (%)**
Original	ResNet-50	97.03 ± 0.74	95.39 ± 0.98	95.78 ± 0.79	95.43 ± 0.84
	VGG16-BN	96.83 ± 0.26	96.39 ± 0.80	95.94 ± 0.78	96.02 ± 0.59
	ViT-B/16	95.33 ± 1.89	95.12 ± 1.73	93.40 ± 2.29	94.07 ± 2.06
With Augmentation	ResNet-50	97.95 ± 0.22	96.98 ± 1.25	97.53 ± 0.70	97.14 ± 0.73
	VGG16-BN	98.03 ± 0.37	97.57 ± 0.67	96.90 ± 0.49	97.15 ± 0.41
	ViT-B/16	97.30 ± 0.49	96.66 ± 1.09	96.50 ± 0.59	96.50 ± 0.79
Δ (Aug. − Orig.)	ResNet-50	+0.93	**+1.60**	+1.75	+1.71
	VGG16-BN	+1.20	+1.17	+0.95	+1.13
	ViT-B/16	**+1.97**	+1.53	**+3.10**	**+2.43**

All three architectures improve on every metric with filtered synthetic augmentation, despite starting from a substantially higher real-only baseline (95–97% accuracy) than on SoySeverityDataset, which reflects the coarser-grained, better-balanced-per-class nature of weed species identification relative to fine-grained disease-severity staging. ViT-B/16 benefits the most, with the largest accuracy gain (+1.97 points), the largest macro-recall gain (+3.10 points), and the largest macro-F1 gain (+2.43 points); ResNet-50 shows the largest macro-precision gain (+1.60 points). ViT-B/16 also has the highest real-only variance across seeds (±1.89 accuracy), which augmentation reduces to ±0.49, the same pattern already observed on SoySeverityDataset, in which the architecture most sensitive to data scarcity benefits most from the additional synthetic coverage.

Table 4.20 reports a paired t-test and a one-sided Wilcoxon signed-rank test on the per-seed accuracy differences, to check whether these gains are attributable to the

augmentation rather than seed variation. The Wilcoxon test, which does not assume normality and is more appropriate given the small sample of five seeds, indicates a consistent, monotonic improvement for all three architectures ($p = 0.0313$, the smallest attainable one-sided p-value at $n = 5$). ViT-B/16 does not reach significance under the paired t-test ($p = 0.1023$) despite having the largest mean gain; this reflects its higher variance rather than an inconsistent effect, since the Wilcoxon result confirms the improvement was directionally consistent across all five seeds.

Table 4.20: Paired t-test and one-sided Wilcoxon signed-rank test results for CottonWeedID15 classification accuracy (original vs. filtered-augmented), $n = 5$ seeds per model.

Model	**Mean Acc. Gain (pp)**	**Paired t-test p**	**Wilcoxon (one-sided) p**
ResNet-50	+0.93	0.0349	0.0313
VGG16-BN	+1.20	0.00044	0.0313
ViT-B/16	+1.97	0.1023	0.0313

Together with the SoySeverityDataset results of Section 4.6.2, the CottonWeedID15 results indicate that the augmentation strategy transfers beyond the primary soybean severity task to a second, independently filtered and class-balanced real-field setting, closing the portability question raised at the start of this section: the benefit of KGDC-Diffusion is a property of the conditioning design, not an artefact of the SoySeverityDataset alone.

4.10 Summary

This chapter presents KGDC-Diffusion, a two-stage generative framework motivated by the scarcity of Early-stage symptoms in field-collected agricultural data. Targeted synthesis can complement field collection by adding quality-controlled images whose distribution overlaps with the real data manifold. The framework pairs a Retrieval-Augmented Generation pipeline, built on an 80-entry expert-curated botanical knowledge corpus, with a depth-only ControlNet conditioned on pseudo-depth

maps estimated automatically from RGB images via zero-shot monocular estimation, requiring no manual spatial annotation. Stage I enriches semantic conditioning with retrieved morphological descriptions, while Stage II encourages coarse geometric realism through the structural layout of real leaves. A realism-score filter, described in Step B of Section 4.4, then selects the synthetic images admitted to the training set.

On the SoySeverityDataset, KGDC-Diffusion achieves an FID of 23.33, a KID of 0.01293, and a Recall of 0.951, outperforming the strongest diffusion baseline, ADM-G (DDPM), by 12.06 FID points. Downstream classifiers augmented with the resulting realism-filtered synthetic images show statistically significant gains across all three tested architectures (ResNet-50, VGG-16, and ViT-B/16), with maximum improvements of 3.76 percentage points in accuracy, 7.11 in macro-precision, 5.87 in macro-recall, and 5.59 in macro-F1. Ablation and sensitivity studies support complementary contributions from retrieval-grounded prompting and depth conditioning; the selected inference settings define the best observed operating region within the evaluated range. External experiments show that the design remains effective after retraining on PlantVillage tomato, CottonWeedID10, DeepWeeds, and CottonWeedID15, rather than demonstrating zero-shot transfer. Runtime measurements were made on an H200, and both training stages were additionally verified on a single NVIDIA L4 with 24 GB VRAM. This demonstrates compatibility with the memory capacity available on high-end 24-GB consumer GPUs without describing either evaluated accelerator as a consumer GPU.

A broader analysis of all proposed methodologies and their interrelationships is presented in the concluding chapter.

Chapter 5

Conclusion and Future Work

This thesis started from the observation developed in Sections 1.1 and 1.2: recognition backbones and generative models have both grown in representational power while leaving the same structural gap open. A shared backbone with independent heads gives one task's prediction no pathway to inform another's features, and a diffusion model conditioned on a class index or a generic text prompt is told *which* class to synthesise but not *what it looks like*. In both cases the conditional structure that governs the domain is missing from the architecture. The two contributions of this thesis are the same move applied at opposite ends of the pipeline: make that conditioning explicit.

5.1 Summary of Contributions

Disease-Conditioned Severity Calibration (DCSC) addresses the recognition gap. Severity is not an absolute visual quantity but a disease-relative one: the evidence for severe damage in one disease is not the evidence for severe damage in another. DCSC conditions the severity pathway on the predicted disease identity through FiLM-based affine modulation, and supervises it with a CORAL ordinal head, a disease-conditioned prototype head, and an Earth Mover's Distance term. On AI Challenger 2018, under a fixed 70/15/15 split with five seeds, DCSC reaches 98.51% disease accuracy and 92.30% severity accuracy with a soft MAE of 0.0909.

The 57% soft-MAE reduction over the VMamba baseline is statistically significant ($p < 0.001$). A representative confusion matrix shows a descriptive 52.84% reduction in serious-to-general errors relative to reproduced TSTC; this count was not itself significance-tested. On the Rice Leaf Severity dataset, where disease accuracy saturates, the soft-MAE reduction reaches 79%, supporting the benefit of the severity pathway independently of the backbone.

Knowledge-Grounded Depth-Controlled Diffusion (KGDC-Diffusion) addresses the generative gap where it matters most, in the early-stage classes that are both the most agronomically useful and the most under-represented. Stage I retrieves expert-validated morphological descriptors from a botanical corpus at inference time, replacing the class-index prompt with a description of the target appearance; Stage II conditions a depth-only ControlNet on pseudo-depth maps from zero-shot monocular estimation, enforcing geometric realism with no manual spatial annotation. On the SoySeverityDataset the framework attains an FID of 23.33, a KID of 0.01293 and a Recall of 0.951, beating the strongest diffusion baseline by 12.06 FID points, and the realism-filtered synthetic images give consistent, statistically significant downstream gains across ResNet-50, VGG-16 and ViT-B/16, up to 3.76 percentage points in accuracy, concentrated in the early-stage classes. Cross-dataset results on PlantVillage tomato, CottonWeedID10, DeepWeeds and CottonWeedID15 show that the design remains effective when retrained for each target dataset; they do not constitute zero-shot generalisation. Supporting both contributions, the **SoySeverityDataset** provides expert-validated soybean imagery across three diseases and three severity levels.

The two results answer the same question. KGDC-Diffusion supplies the conditional information the *data* lacks; DCSC supplies the conditional information the *architecture* lacks. Neither gain came from scaling a backbone or a generator, but from writing an existing conditional dependency into a place where gradients can reach it.

5.2 Limitations

DCSC conditions severity on the *predicted* disease identity, so severity accuracy is structurally bounded by disease accuracy: every disease misclassification supplies the severity pathway with the wrong conditioning vector, and the severity prediction is then made against the wrong disease-specific scale. The disease head in the present design is therefore the ceiling on the severity head, and it was not itself the object of optimisation in this work. Making the disease classifier more robust, particularly under a weaker or open-set taxonomy, is a prerequisite for any further gain in severity accuracy. KGDC-Diffusion depends on a hand-curated botanical corpus, and its retrieval signal is class-level rather than image-level, so a generated image inherits the morphology of its class description rather than of the specimen being augmented; the realism filter is also a post-hoc gate rather than a guarantee built into sampling. Both frameworks were validated offline on curated collections, not under field acquisition conditions.

5.3 Future Work

- **Image-level conditioning.** Replace the hand-written corpus with vision-language-model captions generated per image, raising the conditioning signal from the class level to the instance level and removing the curation bottleneck.
- **Richer cross-task pathways.** FiLM applies a channel-wise affine transform, which is deliberately coarse. Cross-attention between the disease and severity pathways would allow spatially selective conditioning, and infection-type-aware hierarchical priors would let pathologically related diseases share severity structure.
- **A more robust disease head.** Because severity is conditioned on the predicted disease identity, the disease head is the direct route to a higher severity accuracy: every gain there removes a mis-conditioned severity prediction. Hierarchical species-masking, confidence calibration, stronger disease-side augmentation,

and automatic loss weighting in place of fixed multitask weights are all natural candidates, and none of them requires changing the conditioning mechanism itself.

- **Closing the loop.** Use DCSC's per-class severity uncertainty to decide where KGDC-Diffusion should synthesise next, turning augmentation from a fixed preprocessing step into a targeted, uncertainty-driven process.
- **Field deployment.** Agricultural robots collect leaf imagery autonomously but without labels. KGDC-Diffusion offers a route to labelling such streams, while DCSC would need adaptation for on-device real-time severity assessment.

Bibliography

A. Abbas, S. Jain, M. Gour, and S. Vankudothu. Tomato plant disease detection using transfer learning with C-GAN synthetic images. *Computers and Electronics in Agriculture*, 187:106279, 2021. doi: 10.1016/j.compag.2021.106279. URL https://doi.org/10.1016/j.compag.2021.106279.

AI Challenger. AI Challenger 2018: Agricultural disease recognition dataset. https://challenger.ai/dataset/pdr2018, 2018. Plant Disease Recognition Challenge, accessed 2024.

Ali El Bellaj, Aya Benradi, Salman El Youssoufi, Taha El Marzouki, and Mohammed-Amine Cheddadi. Uncertainty-aware ordinal deep learning for cross-dataset diabetic retinopathy grading. *arXiv preprint arXiv:2602.10315*, 2026. doi: 10.48550/arXiv.2602.10315. URL https://arxiv.org/abs/2602.10315.

M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*. OpenReview.net, 2018.

M. Brahimi, M. Arsenovic, S. Laraba, S. Sladojevic, K. Boukhalfa, and A. Moussaoui. Deep learning for plant diseases: Detection and saliency map visualisation. In *Human and Machine Learning*, pages 93–117. Springer International Publishing, 2018.

A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2019.

W. Cao, V. Mirjalili, and S. Raschka. Rank-consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020. doi: 10.1016/j.patrec.2020.11.008. URL https://doi.org/10.1016/j.patrec.2020.11.008.

Rich Caruana. Multitask learning. *Machine Learning*, 1997.

D. Chen, X. Qi, Y. Zheng, Y. Lu, Y. Huang, and Z. Li. Synthetic data augmentation by diffusion probabilistic models to enhance weed recognition. *Computers and Electronics in Agriculture*, 216:108517, 2024. doi: 10.1016/j.compag.2023.108517. URL https://doi.org/10.1016/j.compag.2023.108517.

Dong Chen, Yuzhen Lu, Zhaojian Li, and Sierra Young. Performance evaluation of deep transfer learning on multi-class identification of common weed species in cotton production systems. *Computers and Electronics in Agriculture*, 198:107091, 2022.

J. Chen, D. Zhang, Y. A. Nanehkaran, and D. Li. Detection of rice plant diseases based on deep transfer learning. *Journal of the Science of Food and Agriculture*, 100:3246–3256, 2020.

Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020. doi: 10.48550/arXiv.2009.09796. URL https://arxiv.org/abs/2009.09796.

A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65, 2018. doi: 10.1109/MSP.2017.2765202. URL https://doi.org/10.1109/MSP.2017.2765202.

Boyang Deng and Yuzhen Lu. Weed image augmentation by controlnet-added stable diffusion for multi-class weed detection. *Computers and Electronics in Agriculture*, 232:110123, 2025.

P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In

Advances in Neural Information Processing Systems, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2021.

N. Durasov, T. Bagautdinov, P. Baque, and P. Fua. FiLM-Ensemble: Probabilistic deep learning via feature-wise linear modulation. In *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., 2022.

M. Dutot, L. M. Nelson, and R. C. Tyson. Predicting the spread of postharvest disease in stored fruit, with application to apples. *Postharvest Biology and Technology*, 85: 45–56, 2013.

B. Espejo-Garcia, N. Mylonas, L. Athanasakos, E. Vali, and S. Fountas. Combining generative adversarial networks and agricultural transfer learning for weeds identification. *Biosystems Engineering*, 204:79–89, 2021.

K. P. Ferentinos. Deep learning models for plant disease detection and diagnosis. *Computers and Electronics in Agriculture*, 145:311–318, 2018. doi: 10.1016/j.compag.2018.01.009. URL https://doi.org/10.1016/j.compag.2018.01.009.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059. PMLR, 2016. doi: 10.48550/arXiv.1506.02142. URL https://proceedings.mlr.press/v48/gal16.html.

Zhuangzhi Gao, He Zhao, Zhongli Wu, Yuankai Wang, Gregory Yoke Hong Lip, Alena Shantsila, Eduard Shantsila, and Yalin Zheng. Coral-cvds: a consistent ordinal regression model for cardiovascular diseases grading. In *International Workshop on*

Ophthalmic Medical Image Analysis, Lecture Notes in Computer Science, pages 73–82. Springer, Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72384-1_8. URL https://doi.org/10.1007/978-3-031-72384-1_8.

Ian Goodfellow et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the First Conference on Language Modeling (COLM)*. OpenReview.net, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

S. Gudge, A. Tiwari, M. Ratnaparkhe, and P. Jha. On construction of data preprocessing for real-life soyleaf dataset & disease identification using deep learning models. *Computational Biology and Chemistry*, 117:108417, 2025. doi: 10.1016/j.compbiolchem.2025.108417. URL https://doi.org/10.1016/j.compbiolchem.2025.108417.

V. A. Gulhane and A. A. Gurjar. Detection of diseases on cotton leaves and its possible diagnosis. *International Journal of Image Processing*, 5:590–598, 2011.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 7407–7417. IEEE, oct 2023.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

Le Hou, Chen-Ping Yu, and Dimitris Samaras. Squared Earth Mover's distance-based loss for training deep neural networks. *arXiv preprint arXiv:1611.05916*, 2016. doi: 10.48550/arXiv.1611.05916. URL https://arxiv.org/abs/1611.05916.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708. IEEE, 2017. doi: 10.1109/CVPR.2017.243. URL https://doi.org/10.1109/CVPR.2017.243.

David Hughes and Marcel Salathé. An open access repository of images on plant health to enable the development of mobile disease diagnostics. *arXiv preprint arXiv:1511.08060*, 2015. doi: 10.48550/arXiv.1511.08060. URL https://arxiv.org/abs/1511.08060.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. doi: 10.1109/TBDATA.2019.2921572. URL https://doi.org/10.1109/TBDATA.2019.2921572.

T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119. IEEE, 2020a. doi: 10.1109/CVPR42600.2020.00813. URL https://doi.org/10.1109/CVPR42600.2020.00813.

T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila.

Alias-free generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 852–863. Curran Associates, Inc., 2021.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12104–12114. Curran Associates, Inc., 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/8d30aa96e72440759f74bd2306c1fa3d-Paper.pdf.

A. S. Keceli, A. Kaya, C. Catal, and B. Tekinerdogan. Deep learning-based multi-task prediction system for plant disease and species detection. *Ecological Informatics*, 69:101679, 2022. doi: 10.1016/j.ecoinf.2022.101679. URL https://doi.org/10.1016/j.ecoinf.2022.101679.

Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13893–13902. IEEE, 2020. doi: 10.1109/CVPR42600.2020.01391. URL https://doi.org/10.1109/CVPR42600.2020.01391.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105. Curran Associates, Inc., 2012.

Y. Kurmi, S. Gangwar, V. Chaurasia, and A. Goel. Leaf images classification for the crops diseases detection. *Multimedia Tools and Applications*, 81:8155–8178, 2022.

T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

S. H. Lee, H. Goeau, P. Bonnet, and A. Joly. Conditional multi-task learning for

plant disease identification. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*, pages 3320–3327. IEEE, 2021.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

Q. Liang, S. Xiang, Y. Hu, G. Coppola, D. Zhang, and W. Sun. PD2SE-Net: Computer-assisted plant disease diagnosis and severity estimation network. *Computers and Electronics in Agriculture*, 157:518–529, 2019.

Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and J. Jiao. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., 2024.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00986. URL https://doi.org/10.1109/ICCV48922.2021.00986.

Y. Lu, D. Chen, E. Olaniyi, and Y. Huang. Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review. *Computers and Electronics in Agriculture*, 200:107208, 2022. doi: 10.1016/j.compag.2022.107208. URL https://doi.org/10.1016/j.compag.2022.107208.

Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*,

2018. doi: 10.48550/arXiv.1802.03426. URL https://arxiv.org/abs/1802.03426.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. doi: 10.48550/arXiv.1411.1784. URL https://arxiv.org/abs/1411.1784.

Ishan Misra et al. Cross-stitch networks for multi-task learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4003. IEEE, 2016. doi: 10.1109/CVPR.2016.433. URL https://doi.org/10.1109/CVPR.2016.433.

Sharada P. Mohanty, David P. Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in Plant Science*, 7:1419, 2016. doi: 10.3389/fpls.2016.01419. URL https://doi.org/10.3389/fpls.2016.01419.

Alex Olsen, Dmitry A Konovalov, Bronson Philippa, Peter Ridd, Jake C Wood, Jamie Johns, Wesley Banks, Benjamin Girgenti, Owen Kenny, James Whinney, et al. DeepWeeds: A multiclass weed species image dataset for deep learning. *Scientific Reports*, 9(1):2058, 2019.

Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6887–6896. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00676. URL https://doi.org/10.1109/CVPR52688.2022.00676.

E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. AAAI Press, 2018.

A. Picon, M. Seitz, A. Alvarez-Gila, P. Mohnke, A. Ortiz-Barredo, and J. Echazarra. Crop conditional convolutional neural networks for massive multi-crop plant disease

classification over cell phone acquired images taken on real field conditions. *Computers and Electronics in Agriculture*, 167:105093, 2019. doi: 10.1016/j.compag.2019.105093. URL https://doi.org/10.1016/j.compag.2019.105093.

Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-balancing diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18434–18443. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01768. URL https://arxiv.org/abs/2305.00562.

Alec Radford et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

N. Reimers and I. Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

P. Isaac Ritharson. Severity-based rice disease classification. Kaggle dataset, 2024. URL https://www.kaggle.com/datasets/isaacritharson/severity-based-rice-leaf-diseases-dataset. Accessed 2026.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.org/10.1109/CVPR52688.2022.01042.

H. Sabrol and K. Satish. Tomato plant disease classification in digital images using classification tree. In *International Conference on Communication and Signal Processing*, pages 1242–1246. IEEE, 2016.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

X. Shi, W. Cao, and S. Raschka. Deep neural networks for rank-consistent ordinal regression based on conditional probabilities. arXiv:2111.08851, 2021. URL https://arxiv.org/abs/2111.08851.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2015. URL https://arxiv.org/abs/1409.1556.

Trevor Standley et al. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 9120–9132. PMLR, 2020.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. IEEE, 2016. doi: 10.1109/CVPR.2016.308. URL https://doi.org/10.1109/CVPR.2016.308.

H.-T. Tai, K.-H. Le, and N. L.-T. Nguyen. FormerLeaf: An efficient vision transformer for cassava leaf disease detection. *Computers and Electronics in Agriculture*, 204: 107518, 2023. doi: 10.1016/j.compag.2022.107518. URL https://doi.org/10.1016/j.compag.2022.107518.

A. Tripathi et al. A multitask learning-based vision transformer for plant disease localization and classification. *International Journal of Computational Intelligence Systems*, 17:145, 2024. doi: 10.1007/s44196-024-00517-9. URL https://doi.org/10.1007/s44196-024-00517-9.

Saransh Vashistha. SoySeverityDataset: Severity-Annotated Soybean Leaf Disease

Images for KGDC-Diffusion, 2026. URL https://doi.org/10.5281/zenodo.21180810. [Data set].

Ashish Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017.

D. Wang, J. Wang, Z. Ren, and W. Li. DHBP: A dual-stream hierarchical bilinear pooling model for plant disease multi-task classification. *Computers and Electronics in Agriculture*, 195:106808, 2022a. doi: 10.1016/j.compag.2022.106808. URL https://doi.org/10.1016/j.compag.2022.106808.

F. Wang, Y. Rao, Q. Luo, X. Jin, Z. Jiang, W. Zhang, and S. Li. Practical cucumber leaf disease recognition using improved Swin Transformer and small sample size. *Computers and Electronics in Agriculture*, 199:107163, 2022b. doi: 10.1016/j.compag.2022.107163. URL https://doi.org/10.1016/j.compag.2022.107163.

A. Yang, B. Yang, B. Hui, et al. Qwen2 technical report. arXiv:2407.10671, 2024a. URL https://arxiv.org/abs/2407.10671.

B. Yang, Z. Wang, J. Guo, L. Guo, Q. Liang, Q. Zeng, R. Zhao, J. Wang, and C. Li. Identifying plant disease and severity from leaves: A deep multitask learning framework using triple-branch Swin Transformer and deep supervision. *Computers and Electronics in Agriculture*, 209:107809, 2023. doi: 10.1016/j.compag.2023.107809. URL https://doi.org/10.1016/j.compag.2023.107809.

B. Yang, M. Li, F. Li, Y. Wang, Q. Liang, R. Zhao, C. Li, and J. Wang. A novel plant type, leaf disease and severity identification framework using CNN and transformer with multi-label method. *Scientific Reports*, 14:11664, 2024b. doi: 10.1038/s41598-024-62452-x. URL https://doi.org/10.1038/s41598-024-62452-x.

G. Yang, Y. He, Y. Yang, and B. Xu. Fine-grained image classification for crop disease based on attention mechanism. *Frontiers in Plant Science*, 11:600854, 2020. doi: 10.3389/fpls.2020.600854. URL https://doi.org/10.3389/fpls.2020.600854.

L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10371–10381. IEEE, 2024c.

L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847. IEEE, 2023.